National Agency for Language Development and Cultivation
Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia

ASIALEX
The Asian Association for Lexicography

ASIALEX 2021

June 12—14
Jakarta, Indonesia

PROCEEDINGS

of the 14th International Conference of the Asian Association for Lexicography

"Lexicography and Language Documentation"

Editors
Dora Amalia
Azhari Dasman Darnis
Amat Triatna
Dewi Khairiah

ASIALEX
The Asian Association for Lexicography

ASIALEX
2021
June 12—14
Jakarta, Indonesia

PROCEEDINGS

of the 14th International Conference of the Asian Association for Lexicography

“Lexicography and Language Documentation”

Editors
Dora Amalia
Azhari Dasman Darnis
Amat Triatna
Dewi Khairiah

National Agency for Language Development and Cultivation
Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia

Jakarta
2021

The Asian Association for Lexicography (ASIALEX) Proceedings Book

ISBN: 978-623-307-134-5

EDITORS
Dora Amalia
Azhari Dasman Darnis
Amat Triatna
Dewi Khairiah

COVER DESIGN
Nurjaman

DATE OF PUBLISHING
August 1, 2021

PUBLISHER
National Agency for Language Development and Cultivation
Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia
Jl. Daksinapati Barat 4 No.11, RT.11/RW.14 Rawamangun, Kec. Pulo Gadung,
Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, 13220
Phone.: (021) 4750406
Email: badanbahasa@kemdikbud.go.id

The Asian Association for Lexicography (ASIALEX)
Proceedings Book
"Lexicography and Language Documentation"

BOARD OF ASIALEX 2019—2021

President:	Vincent Ooi (National University of Singapore, Singapore)
Vice-President:	Hai Xu (Guangdong University of Foreign Studies, People's Republic of China)
Secretary:	Ai Inoue (National Defense Academy, Japan)
Treasurer:	Kilim Nam (Kyungpook National University, Korea)

BOARD MEMBER

Ferdi Bozkurt (Anadolu University, Turkey)
Jesus Federico (Tuting) Hernandez (The Hong Kong Polytechnic University, Hong Kong)
Cuilian Zhao (Sichuan International Studies University, People's Republic of China)

CONVENER OF ASIALEX 2021

Dora Amalia (National Agency for Language Development and Cultivation, Indonesia)

CONVENER OF ASIALEX 2022

Anmin Wang (Guangxi University for Nationalities, People's Republic of China)

CO-CHIEF EDITOR OF LEXICOGRAPHY

Shigeru Yamada (Waseda University, Japan)

PAST PRESIDENT

Rachel Edita O. Roxas (National University, The Philippines)

ORGANISING COMMITTEE

Chair

Dr. Dora Amalia

Co-Chair

Dr. Dewi Puspita

Members

Azhari Dasman Darnis
Vita Luthfia Urfa
Dina Alfiyanti Fasa
Winda Luthfita
Selly Rizki Yanita
Nikita Daning Pratami
Vita Muflihah Fitriyani
Raymond Allan
Dewi Khairiah
Amat Triatna
Dzien Nuen Almisri
Tita Nurajeng Miyasari
Rissa Yulia Pungkysari
Rinda Yosa
Ivana Maliha
Meryna Afrila
Tri Indira Satya P.
Nurjaman
M. Irfan Riansyah
Devi Virhana
M. Cesar Nurkarim
Dira Hildayani
Kunkun Purwati
Endah Nur Fatimah

Steering Committee and Reviewer

1. Ai Inoue, National Defense Academy, Japan
2. Amy Chi, Hong Kong University of Science and Technology, Hong Kong SAR
3. Anmin Wang, Guangxi University for Nationalities, PRC
4. Arleta Adamska-Salaciak, Adam Mickiewicz University, Poland
5. Dora Amalia, National Agency for Language Development and Cultivation, Indonesia
6. Gilles-Maurice De Schryver, Ghent University, Belgium; University of Pretoria, South Africa
7. Hai Xu, Guangdong University of Foreign Studies, Guangzhou, PRC
8. Ilan Kernerman, K Dictionaries Ltd, Israel
9. Jesus Federico Hernandez, University of the Philippines Diliman, The Philippines
10. Kilim Nam, Kyungpook National University, South Korea
11. Li Lan, Chinese University of Hong Kong, Shenzhen
12. Totok Suhardijanto, Universitas Indonesia, Indonesia
13. Vincent Ooi, National University of Singapore, Singapore
14. Yongwei Gao, Fudan University, PRC
15. Yukio Tono, Tokyo University of Foreign Studies, Japan

WELCOME REMARKS

Welcome to the first virtual conference of Asialex. It is an honor for us to be the host of The 14th International Conference of The Asian Association for Lexicography (Asialex) which has finally been held, although, unfortunately, it is virtual. We all know some situations make this all happen.

This conference was supposed to be held in 2020 but had to be postponed twice until it could finally be held now. The conference took the theme "Lexicography and Language Documentation". The topic is very relevant to the linguistic situation in Indonesia because it is a country with a language diversity. According to language mapping held by Badan Bahasa, there are 718 regional languages that are not well documented, so their extinction leaves no trace. The discussion results from this conference will be our reference in dealing with local language documentation issues.

There are 4 keynote speakers who will present their papers. They are E. Aminudin Aziz from the National Agency for Language Development and Cultivation or Badan Bahasa, René van den Berg (SIL International), Rufus Gouws (Stellenbosch University), and Li Lan (The Chinese University of Hong Kong, Shenzen).

Since it was announced in October 2020, there have been 64 abstracts submitted and 56 were accepted, but only 42 papers will be presented at this conference with 54 presenters attending. In addition, there are 13 non-presenter participants, 5 interpreters, and the committee.

One of the highlights of this conference is the provision of Indonesian and English interpreters. This is our effort to provide opportunities for Indonesian speakers who will present their papers in Indonesian so that foreign participants can understand them or vice versa. Supported by 5 professional interpreters, I hope this conference can run smoothly and be understood by all participants.

I would like to thank the Board Members of Asialex for supporting us to hold this conference virtually. At first, we proposed a hybrid format, face-to-face and virtual. But looking at the evolving conditions and for the safety of all participants, we finally decided to hold a fully virtual conference.

Finally, I hope this conference can become an information and knowledge exchange forum about the state-of-the-art of Asian lexicography.

Enjoy the conference and thank you.

Dr. Dora Amalia

Convener of Asialex 2021

CONTENTS

THE USE OF LEXICAL BUNDLES IN ONLINE INDONESIAN COMPREHENSIVE DICTIONARY (KBBI DARING)

Adi Budiwiyanto, Totok Suhardijanto
Faculty of Humanities, Universitas Indonesia
adibudiwiyanto@gmail.com; suhardiyanto@gmail.com

Abstract

Research on lexical bundles in the last few decades has been focusing mostly on written registers, especially academic writing. In this study, we investigate the use of lexical bundles on different genre: dictionary. As lexical bundle is a formulaic language which is specific to a particular register, we hypothesize that there are particular lexical bundles used in dictionary. The research question of this study focuses on the extent to which lexical bundles are used in KBBI Daring, especially on the lemma, definition, and example section. This study used a corpus-based approach. The lexical bundles used as reference bundles are 420 lexical bundles extracted from IndonesianWeb Corpus (SketchEngine). The bundles are then analyzed for their use in KBBI Daring in terms of their frequency, structure, and fuction. The results showed that the use of lexical bundles in KBBI Daring was mostly found in the definition section. The bundles found were generally in the form of phrase rather than clause. In terms of structure, lexical bundles are dominated by incomplete structures. The bundles, either in the definition or example section, were mostly in the pattern of *yang*-clause fragment, such as *yang digunakan untuk*, *yang terdiri atas*, *yang terbuat dari*, *yang berasal dari*, and *yang berhubungan dengan*, that have descriptive function. This study also found a number of potential lexical bundles for KBBI, such as *oleh karena/sebab itu*, *di samping itu*, *dengan kata lain*, *dalam hal ini*, and *di sisi lain*. Therefore, it is suggested to include them as sublemmas and arrange them based on their core elements: *karena*, *sebab*, *samping*, *kata*, *hal*, and *sisi*.

Keywords: lexical bundles, dictionary, KBBI Daring, corpus-based, frequency

1 INTRODUCTION

Lexical bundles have received increasing attention over the last three decades. Altenberg (1999) used a frequency-based approach to examine recurrent word combinations in spoken English. Biber et al. (1999) studied extended collocations, called lexical bundles, in four registers, namely conversation, academic prose, fiction, and news. Biber (2006) also developed a study on the use of lexical bundles in spoken and written registers at university. Hyland (2008) focused on lexical bundle studies on theses and dissertations in four different disciplines. Salazar (2014) also examines lexical bundles in native and non-native scientific writing. Jalilifar et al. (2017) and Kwary et al. (2017) identified English lexical bundles in interdisciplinary journal articles.

In Indonesian, there has not been many studies of lexical bundles that have been carried out. Novita and Kwary (2018) examined Indonesian lexical bundles to compare them to the English translations produced by student translators and professional translators. Samodra and Pratiwi studied Indonesian lexical bundles to compare to the English translations in undergraduate abstracts. Meanwhile, Budiwiyanto and Suhardijanto (2019, 2020) put more attention to the use of Indonesian lexical bundles in written academic discourse on legal studies and research articles from several disciplines.

The term *lexical bundle* was first used by Biber et al. (1999) in *Longman Grammar of Spoken and Written English*. Biber et al. (1999) defines lexical bundles as recurring sequences of three or more words, regardless of their idiomaticity, and regardless of their structural status. Lexical bundles are simply sequence of words that commonly go together in natural discourse. The bundles are identified by a frequency and range threshold. The frequency threshold indicates that the lexical bundles do not occur accidentally, while the range threshold indicates that the lexical bundles are not an idiosyncratic use of the individual speaker or writer.

Lexical bundles have been classified in terms of their structures as well as their functions (Biber & Barbieri, 2007; Byrd and Coxhead, 2010; Conrad & Biber, 2004; Cortes, 2004; Hyland, 2008; Salazar, 2014). Relating to their structure, only 15 percent of lexical bundles in conversation can be regarded as complete phrases or clauses, while less than 5 percent of the lexical bundles in an academic prose represent complete structural units. Moreover, almost all the bundles bridge two structural units and are mostly not idiomatic (Biber, 2006).

Biber et al. (1999, 2006), Hyland (2008), Byrd and Coxhead (2010) functionally classified the lexical bundles into three category with different labels. However, they are essentially similar. Biber divided the function into referential bundles, discourse organizers, and stance expressions; Hyland classified them into reserach-oriented bundles, text-oriented bundles, and participant-oriented bundles; while Byrd and Coxhead used presenting content, text organizer, and expressing stance. The functions in their taxonomy refer to the meanings and purposes of the language. The functions try to organize the discourse according to situations or contexts.

Dictionary describes the vocabulary of a language. It tells its reader s the ways in which that words typically contributes to the meaning of an utterance, the ways in which it combines with other words, and the types of text that it tends to occur in (Atkins & Rundell, 2008, p. 45). Clearly it is desirable that the description given in a dictionary is reliable. According to Atkins & Rundell, a reliable dictionary is one whose generalization about word behaviour approximate closely to the ways in which people normally use and understand language when engaging in real communicative acts. To achieve this goal, one way is to use vocabulary that has a high frequency and is widely used in society (p.48). This statement implies that the headwords, definitions, and examples in dictionary, need to take this consideration as well.

Word combination is one of the elements that must exist in a dictionary entry. Word combinations are central to part of the vocabulary of most languages, and need to be accounted for in the dictionary. They are particularly important for learners' dictionaries, both monolingual and bilingual, since language learner may not recognize them as significant units of meaning, cannot usually compose them, and will often have problems understanding them. Some may be easy to spot, but many are less idiomatically salient (Atkins & Rundell, 2008, p. 167).

According to Atkins and Rundell (p. 407), definitions exist to catalogue the meanings in a language, but their practical purpose is to resolve the communicative needs of dictionary. Dictionary succeed when they get two things right: content and form. The precise configuration will be determined by the needs and skills of the users. Meanwhile, example sentences are vital component. Their function is to support and illustrate every linguistic fact and as a source of data from which lexicographers construct their entries. The nature of examples will vary according to the type of dictionary and needs and expectations of its users. However, at least examples should be natural and typical, informative, and intelligible. Typicality is easy enough to recognize: a large corpus will show the contexts, syntactic patterns, collocations, and multiword expressions in which a word is most frequently found, and these represent its typical forms of ehaviour. Naturalness is a more intuitive and less objective measure. Reccurence is important here to see a text or utterance is natural or not. A natural example is one that maintains a consistent register.

The dictionary studied in this research is the *Kamus Besar Bahasa Indonesia Daring* (*Online Indonesian Comprehensive Dictionary* [KBBI Daring]). KBBI Daring is an online version of printed KBBI

Fifth Edition[1] (2016). Beside printed and online version, KBBI also published offline version and braille version. KBBI Daring is a proscriptive[2] monolingual dictionary that was first launched in 28 October 2016 by the Ministry of Education and Culture of Indonesia. It was compiled by lexicografers at Badan Pengembangan dan Pembinaan Bahasa (National Agency for Language Development and Cultivation). This online version involves communities in contributing, compiling and editing the entries. Presently, KBBI Daring has 114,665 entries, comprising 51,874 headwords, 27,280 derrivations, 31,945 word combinations, 2,075 proverbs, 270 idioms and 1,132 foreign and local expressions. Totally there are 133,709 meaning and 30,480 examples (https://kbbi.kemdikbud.go.id/Beranda/Statistik accessed May 3, 2021).

This study seeks to fill in the gap of the research of lexical bundles by focusing on different register, namely dictionary. As Biber said that lexical bundle is a formulaic language which is specific to a particular register, we hypothesize that there are particular lexical bundles used in dictionary. The research questions of this study focus on the extent to which lexical bundles are used in KBBI Daring, especially on the lemma, definition, and example section.

2 METHOD

This study used corpus-based approach. The corpus used in this study is taken from IndonesianWeb Corpus (SketchEngine) that consists of 109,236,814 word tokens with 27,051 documents. The lexical bundles extracted from this corpus by N-Grams and are used as reference bundles. The extracted bundles must occur at least 10 times per million words. This threshold followed the criteria set by Biber (1999). The extraction yielded 478 bundles. The bundles, then, were filtered and sorted to avoid foreign words and proper names. The result was 420 bundles that were analysed for their use in KBBI Daring. Here are the top 50 of the bundles.

Table 1 Top 50 most frequent bundles in SketchEngine

No	Bundle	Frequency	Set	No	Bundle	Frequency	Set
1	yang ada di	15162	3	26	pada waktu itu	4216	3
2	oleh karena itu	14101	3	27	menjadi salah satu	4132	3
3	dalam hal ini	9284	3	28	tahun yang lalu	4125	3
4	yang dilakukan oleh	7633	3	29	apa yang telah	4124	3
5	yang berasal dari	7131	3	30	sampai saat ini	3960	3
6	merupakan salah satu	7036	3	31	dengan cara yang	3844	3
7	yang berada di	6848	3	32	apa yang terjadi	3841	3
8	sama sekali tidak	6426	3	33	yang tidak dapat	3822	3
9	yang lebih baik	6389	3	34	di samping itu	3802	3
10	yang luar biasa	5764	3	35	dengan apa yang	3774	3
11	yang berkaitan dengan	5619	3	36	yang lebih tinggi	3539	3
12	tidak ada yang	5546	3	37	yang berhubungan dengan	3536	3
13	yang selama ini	5352	3	38	yang akan dating	3529	3
14	adalah salah satu	5341	3	39	salah satu dari	3497	3
15	yang terjadi di	5237	3	40	di luar negeri	3472	3
16	pada saat itu	5190	3	41	yang dimaksud dengan	3388	3

[1] The first edition was published in 1988; the second edition was in 1991; the third edition was in 2000; and the fourth edition was in 2008.

[2] not only describing or forbidding, but making recommendations in cases of variation

17	dengan kata lain	5131	3	42	di antara mereka	3383	3
18	sebagai salah satu	5000	3	43	tidak akan pernah	3366	3
19	di seluruh dunia	4996	3	44	sumber daya alam	3324	3
20	satu sama lain	4649	3	45	yang ada dalam	3323	3
21	di dunia ini	4583	3	46	bagi mereka yang	3322	3
22	yang terdiri dari	4537	3	47	apa saja yang	3318	3
23	yang lebih besar	4501	3	48	di sisi lain	3274	3
24	masuk ke dalam	4473	3	49	orang yang tidak	3200	3
25	hak asasi manusia	4274	3	50	yang tinggal di	3188	3

Next, each bundle was examined to know whether or not it was used in every section of lemma, definition and example. The bundles, then, were analysed in terms of frequency, structure and function. The analysis of function used in this study adopted the classification of Hyland (2008) and Byrd and Coxhead (2010). Hyland classiffied the function into three categories: research-oriented, text-oriented, and participant-oriented. Hyland's taxonomy reflects the three major metafunctions of language, ideational, textual, and interpersonal.

The research-oriented bundles function to help structure experience and activity of real world, consisting of five subfunction: 1) location, indicating time/place, e.g. *at the same time*, *at the beginning* of; 2) procedure, e.g. *the use of the*, *the purpose of the*; 3) quantification, e.g. *a wide range of*, *one of the most*; 4) description, e.g. *the structure of the*, *the size of the*; 5) topic, e.g. *in the United States*, *the currency board system*. The text-oriented bundles deal with meaning of text and its organization, comprising 1) transition signals, establishing additive or contrastive links between elements, e.g. *on the other hand*, *in addition to the*; 2) resultative signals, marking inferential or causative relations between elements, e.g. *as a result of*, *it was found that*; 3) structuring signals, being text-reflexive markers which organize stretches of discourse or direct the reader elsewhere in text, e.g. *in the present study*, *in the next section*; 4) framing signals, situating arguments by specifying limiting conditions, e.g. *in the case of*, *on the basis of*. The participant-oriented focuses on the writer or the reader, consistinf of 1) stance, conveying the writer's attitudes and evaluations, e.g. *may be due to*, *it is possible that*, and 2) engagement, addressing readers directly, e.g. *as can be seen*. Hyland's taxonomy reflects the three major metafunctions of language, ideational, textual, and interpersonal. However, since this register is not related to academic writing, we did not use the terminology indroduced by Hyland, but by Byrd and Coxhead, namely *presentation of content*, *organization of discourse/text* and *expression of attitude by the writer/speaker*. In our opinion, the terminology used by Byrd and Coxhead is easier to understand for learners or laymen.

3 RESULTS AND DISCUSSION

Of the 420 lexical bundles used as reference bundles, there are 357 lexical bundles that have been used in KBBI Daring. The bundles are spread in the lemmas, definitions, and examples section. In the lemma section, there are only 4 variations of lexical bundles with an occurrence rate of 0.04%, namely *hak asasi manusia, sumber daya manusia, sumber daya alam,* and *dewan perwakilan rakyat*. In the definition section, there are 312 variations of the lexical bundle used with the occurrence rate of 89.76%, among which are *yang digunakan untuk, yang terdiri atas,* and *yang terbuat dari*, while in the example section, there are 275 lexical bundles with the occurrence rate of 10.2%, for example *tidak ada yang, yang ada di, orang yang tidak*. The frequent use of lexical bundles in the definition section seems reasonable because the definition section is the main element in lexicography. Different from the definition, the example elements are supporting and their presence can be provided when needed. In the example section, although the lexical bundle variation is quite high, the frequency of occurrence is low.

Table 2 Top 10 lexical bundles in SketchEngine which are not used by KBBI Daring

No	**Bundle**	**Freq.**	**Set**
1	oleh karena itu	14101	3
2	yang terdiri dari	4537	3
3	yang dimaksud dengan	3388	3
4	oleh sebab itu	3139	3
5	sebagaimana dimaksud pada	2918	3
6	dimaksud pada ayat	2895	3
7	sebagaimana dimaksud pada ayat	2834	4
8	sebagaimana dimaksud dalam	2831	3
9	apa yang kita	2509	3
10	adalah sebagai berikut	2489	3

Meanwhile, there are 63 lexical bundles that are not used in the three sections (see Table 2), consisting of 61 three-word bundles and 2 four-word bundles. These lexical bundles are generally in the form of phrases (36 bundles) and these bundles are generally incomplete structures (46 bundles). In terms of function, these lexical bundles are generally used to present contents (36 bundles) and organize texts (21 bundles). Although the bundles do not appear in KBBI Daring, they are indeed potential for KBBI, both as lemmas, supporting definitions, and supporting examples. In our opinion, of 63 lexical bundles which are not used, there are at least nine potential bundles to be inculed as lemma, namely *oleh karena itu, oleh sebab itu, sejak saat itu, selain itu juga, tak lama kemudian, di lain pihak, dan dengan demikian, beberapa waktu lalu,* and *begitu juga dengan.* These bundles have their own functions. The bundles *oleh karena itu*, *oleh sebab itu*, and *dan dengan demikian* are used to introduce the logical result of something that has just been mentioned. The bundles *selain itu juga* and gugus *begitu juga dengan* are used to introduce additive information. The bundles *sejak saat itu* and *beberapa waktu lalu* are used to indicate the time of an event. The bundle *tak lama kemudian* is used to express time sequences. The last, the bundle *di lain pihak* is used to introduce different points of view, ideas, etc., especially when they are opposites. The following are the examples of their usage.

a. Metode ini dianggap metode kontrasepsi yang permanen dan nonreversible. Oleh karena itu, metode ini tidak direkomendasikan bagi dewasa muda.
b. Yang menarik, dalam komunikasi di dunia maya, ada gejala baru, yaitu seseorang sengaja menutup identitasnya, *dan dengan demikian* menghindarkan diri dari sebuah tanggung-jawab sosial.
c. *Selain itu juga* dibahas mengenai rencana pembentukan komisi bersama untuk meningkatkan kerjasama di bidang ekonomi, perdagangan, dan iptek.
d. *Begitu juga dengan* pertanyaan-pertanyaan yang dilontarkan oleh para peserta bedah buku ini.
e. *Sejak saat itu*, semua orang yang aku kenal tak mau bicara dengan diriku.
f. *Beberapa waktu lalu* Obama sempat memberikan pernyataan akan meyakinkan umat muslim bahwa Amerikan bukanlah musuh.
g. *Tak lama kemudian* terdengar suara ketukan di pintu.
h. Mereka tidak suka perubahan. *Di lain pihak*, banyak orang AS suka perubahan.

Based on the core element, phrasal bundle can be grouped into: nominal bundles, verbal bundles, adjective bundles, prepositional bundles, and adverbial bundles. The nominal bundle can be divided inti proper name bundles, terminological bundles, and binomial bundles. Meanwhile, clausal bundle can be divided into free-clause bundles and bound-clause bundles. The bound-clause bundles are subdivided into subordinate-clause bundles and relative-clause bundles. These groupings can be illustrated in the following figure.

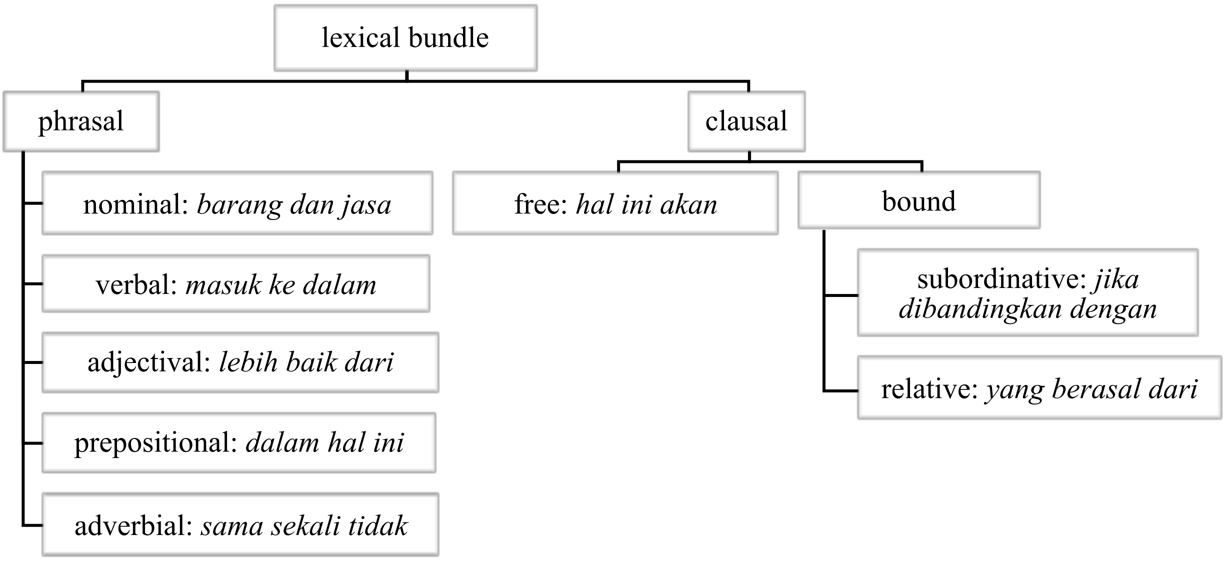

lexical bundle
phrasal
clausal
nominal: *barang dan jasa*
verbal: *masuk ke dalam*
adjectival: *lebih baik dari*
prepositional: *dalam hal ini*
adverbial: *sama sekali tidak*
free: *hal ini akan*
bound
subordinative: *jika dibandingkan dengan*
relative: *yang berasal dari*

Figure 1 Classification of lexical bundles used in KBBI Daring

In terms of function, the varied lexical bundles in KBBI Daring are mostly found in presenting-context bundles which amount to 82.4%, followed by expressing-attitude bundles by 10.1%, and organizing-text bundles by 7.6%. In presenting-content-bundles, the description function has the highest bundle variation, while the lowest is the topic function. In organizing-text bundles, transition function has the most dominant variation among the other two functions. Meanwhile, stance function is the only function that appears in expressing-attitude bundles and has a fairly high variety of bundles. However, overall the description function is the function with the highest, while the framing function is the lowest in terms of variation.

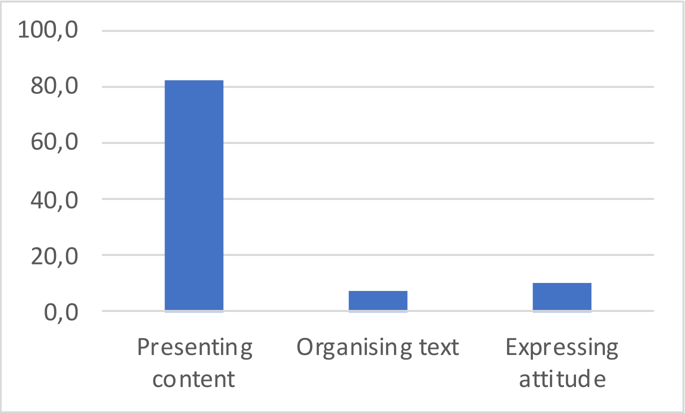


Figure 2 Functional distribution of lexical bundles in KBBI Daring

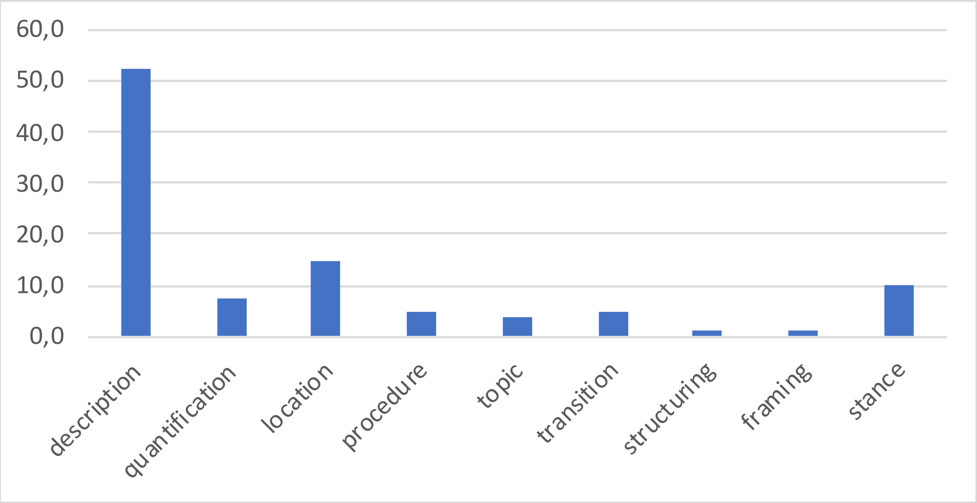


Figure 3 Subfunctional distribution of lexical bundles in KBBI Daring

In addition, this study found that of the 357 bundles existing in KBBI, there were 3 common lexical bundles, namely the bundles that appear together in the three sections: lemma, definition, and example. The bundles include *hak asasi manusia*, *sumber daya alam*, and *dewan perwakilan rakyat*. These three bundles are terminological bundles and belong to topic subfunction.

After explaining how the overall use of lexical bundles in the KBBI, the following subsection consecutively describes how the lexical bundle is used within three parts, namely entries, definitions, and examples, in terms of frequency, structure, and function.

3.1 The use of lexical bundles in lemma section of KBBI

In the entry section, as previously mentioned, there are four variations of the lexical bundles, namely *hak asasi manusia, sumber daya alam, sumber daya manusia,* and *dewan perwakilan rakyat*. These bundles are all complete in structure and nominal-phrase bundles. In terms of function, the bundles belong to topic function. in the form of terms that are usually used in certain registers. The use of the lexical bundle in KBBI can be seen below.

a. **hak**^{1}» **hak asasi manusia** hak yang dilindungi secara internasional (yaitu deklarasi **PBB Declaration of Human Rights), seperti hak untuk hidup, hak kemerdekaan, hak untuk** memiliki, hak untuk mengeluarkan pendapat
b. **sumber** » **sumber daya alam** potensi alam yang dapat dikembangkan untuk proses produksi
c. **sumber** » **sumber daya manusia** potensi manusia yang dapat dikembangkan untuk proses produksi
d. **dewan** » **Dewan Perwakilan Rakyat** badan legislatif yang anggotanya terdiri atas para wakil rakyat yang dipilih baik secara langsung maupun tidak langsung, bertugas membuat undang-undang dan menetapkan anggaran pendapatan dan biaya negara; parlemen

Although only four bundles are used in the KBBI, there are several lexical bundles used in the definitions and examples sections that have the potential to be included into the dictionary as lemmas.

These bundles are as follows:

dalam hal ini
dalam jangka waktu
dalam kurun waktu
dengan kata lain
di samping itu

di muka bumi
di seluruh dunia
di bawah pimpinan
di satu sisi
di sisi lain

hingga saat ini
sampai saat ini
kedua belah pihak
maka dari itu
pada saat yang sama

pada waktu itu
sama sekali tidak
satu sama lain
sedikit demi sedikit
tersebut di atas.

These lexical bundles have particular functions in the text, either as a tool for conveying ideas, for organizing ideas or for expressing attitude. For example, the bundles *dalam hal ini* and gugus *pada saat yang sama* can be used as framing, i.e. situating arguments by specifying limiting conditions. Bundels *di muka bumi* dan *di seluruh dunia*—which are used to denote location of place—are perhaps easy to understand for native speakers. However, they may be difficult for non-native speakers. The bundle *di muka bumi* is a fixed expression. If the word *muka* is replaced by the word *depan*, which is its synonym, so it becomes *di depan bumi*, the meaning of these two bundles are absolutely different. Likewise, the word *seluruh* in the bundle *di seluruh dunia* will have a different meaning if it is replaced by the word *semua*.

3.2 The use of lexical bundles in definition section

In the definition section, there are 312 lexical bundles that have been used, consisting of 310 three-word bundles, such as *yang digunakan untuk*, *yang terdiri atas*, dan *yang terbuat dari*, and 2 four-word bundles, namely *ilmu pengetahuan dan teknologi* and *masa yang akan datang*. The bundles used in this section tend to be in the form of a phrase (51.6%), while clausal bundles are also quite high (48.6%). The structure of these bundles, either phrasal or clausal, is generally incomplete (64.7%). In the phrasal bundles, the preposition-based bundles (28.5%) with the pattern *Prep + NP* (*fragment*) were most frequent occured. Meanwhile, in the clausal bundles, the bound-clause bundles with the pattern *yang + passive verb + prepositional-phrase fragment* are the most frequently used (see Table 3). The following are some examples in use.

a. **al.ki.sah** n ungkapan *yang digunakan untuk* memulai sebuah cerita atau hikayat
b. **da.ging** n **1** gumpal (berkas) lembut *yang terdiri atas* urat-urat pada tubuh manusia atau binatang (di antara kulit dan tulang);
c. **ben.drat** n Jw tali *yang terbuat dari* besi baja, ukurannya relatif kecil, fungsinya sebagai pengikat besi dengan besi dan sebagainya
d. **ke.bu.mi.an** *n* hal *yang berhubungan dengan* bumi
e. **ga.lur** n Tern ciri khas *yang terdapat pada* sekelompok ternak dalam satu bangsa yang ada pada ternak lain dalam bangsa yang sama

Table 3 Top 10 lexical bundles in defintion section

No	**Bundle**	**Freq. KBBI**	**Freq. SE**	**Set**
1	yang digunakan untuk	597	3039	3
2	yang terdiri atas	464	1089	3
3	yang terbuat dari	400	1529	3
4	yang berasal dari	345	7131	3
5	yang berhubungan dengan	284	3536	3
6	yang terletak di	187	2904	3
7	yang hidup di	187	1624	3
8	yang terdapat di	186	2213	3
9	yang berkaitan dengan	172	5619	3
10	yang terdapat pada	170	1127	3

In terms of function, there are eight functions used in the definition section which are spread into three groups: presenting content, organizing text, and expressing attitude. Among the three main groups, the presenting-content bundles are the highest in terms of occurrence (96.65%) as well as variation (84.62%). Then, it was followed by the organizing-text bundles (8.01%) and at last the expressing attitude bundles (7.37%) in terms of frequency of occurrence. However, in terms of variation, the expressing-attitude bundles (2.17%) are slightly higher than the organizing-text bundles (1.18%), as can be seen in Figure 4.

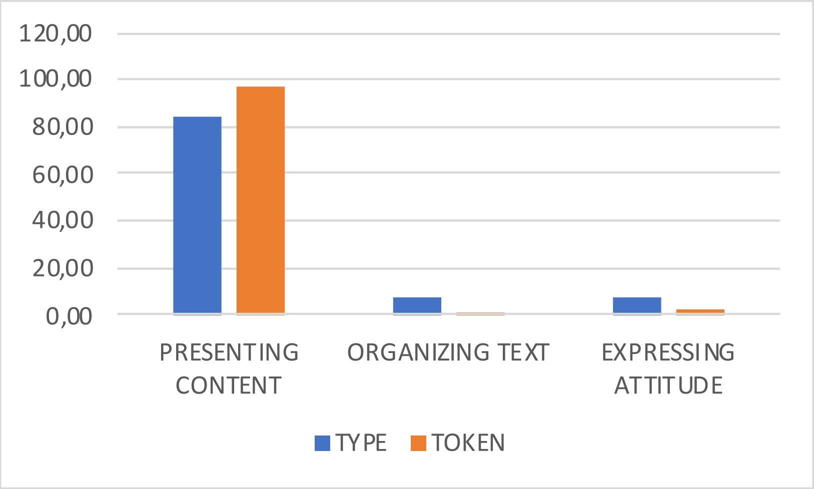


Figure 4 The functional distribution of lexical bundles in definition section

In the presenting-content bundles, the bundles with the description function are the highest in their occurrence, and followed by the location function. The procedure and quantification functions are slightly different in number of occurrence. Meanwhile, both framing and structuring functions are the least used bundles. In the organizing-text bundles, transitional bundles have the highest level of usage. The details can be seen in following figure.

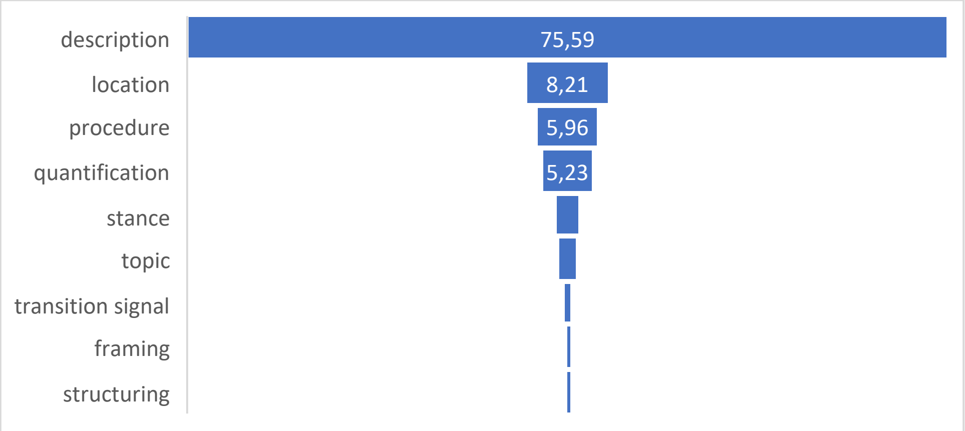


Figure 5 The function of lexical bundlesin definition section based on occurrence

In terms of variation, the bundles with the description function are the most varied, followed by the location function. Subsequently, bundles with functions of establishment, quantification, transition, and procedures are respectively in the middle position. Meanwhile, framing function is the lowest level of variation. In the organizing-text bundles, as with their frequency of occurrence, the transition bundles also have the highest level of variation. The ranking of the eight sub-functions can be seen below.

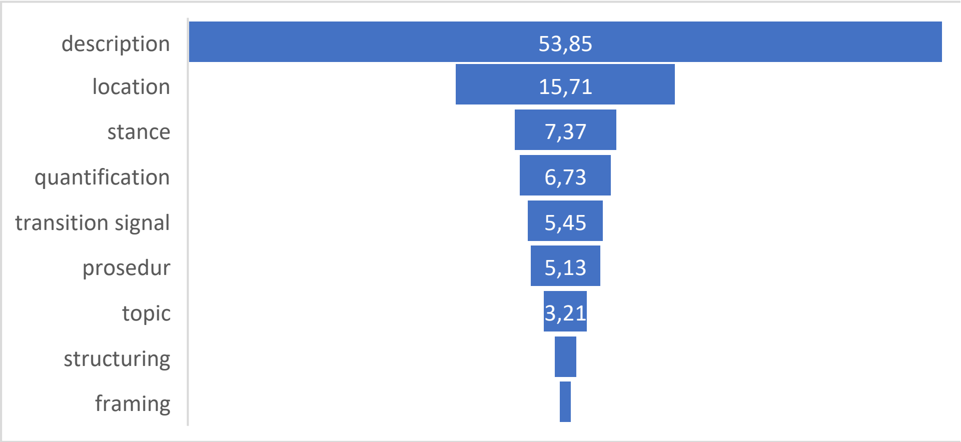


Figure 6 The function of lexical bundles in definition section based on variation

3.3 The use of lexical bundles in example section

In the example section there are 275 lexical bundles used, which consist of 272 three-word bundles, such as *yang ada di*, *masuk ke dalam*, and 3 four-word bundles, namely *ilmu pengetahuan dan teknologi*, *masa yang akan datang*, and *tuhan yang maha esa* (see Table 5). The lexical bundles in this example section are generally incomplete structure. The use of the phrasal bundles (49.45%) is almost equivalent to the clausal bundles (50.55%), only a difference of 1.1%. In the phrasal bundles, the preposition-based bundles (24.36%) with the pattern *Prep + NP* (*fragment*) was the most frequently appeared. Meanwhile, In the clausal bundle level, the bound-clausal bundles with pattern of *yang + intransitive verb + prepositional phrase fragment* and *yang + passive verb + prepositional phrase fragment* are two most frequently used. Here are some bundles in use.

a. semua jabatan terisi, *tidak ada yang* lowong
b. segala *yang ada di* dunia fana belaka
c. perbuatannya seperti kelakuan *orang yang tidak* beradab
d. karena sifatnya yang buruk itu, *banyak orang yang* membencinya
e. faktor adanya kesempatan *merupakan salah satu* pemengaruh perilaku menyontek

Table 5 Top 10 lexical bundles in example section of KBBI

No	Bundle	Freq. KBBI	Freq. SE	Set
1	tidak ada yang	32	5546	3
2	yang ada di	25	15162	3
3	orang yang tidak	20	3200	3
4	banyak orang yang	20	3050	3
5	masuk ke dalam	19	4473	3
6	ke luar negeri	19	1808	3
7	di rumah sakit	17	2162	3
8	tahun yang lalu	17	4125	3
9	dalam bahasa indonesia	15	1688	3
10	merupakan salah satu	15	7036	3

Among the three main groups, the presenting content bundle is the group with the highest rate in terms of occurrence, by 81.45%, and in terms of variation, by 78.9%. Then, it was followed by the expressing attitude bundle, by 15.89% in terms of occurrence, and 12.36% in terms of variation. The last is the organizing text bundle with 5.1% in terms of frequency of use and 6.18% in terms of variation. The comparison of the three main functions of the lexical bundles can be seen in the following figure.

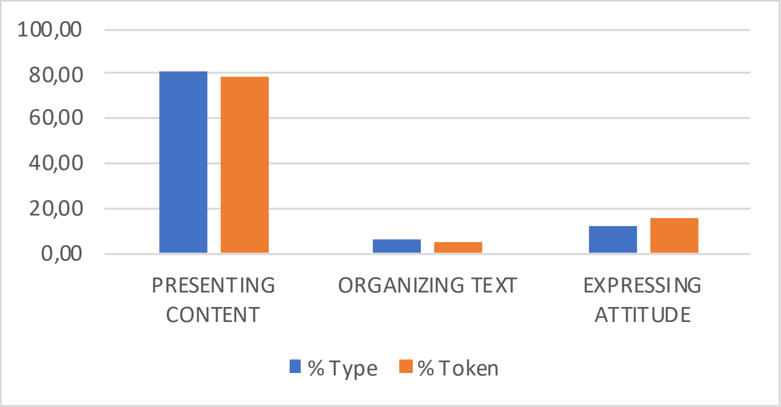


Figure 7 The function of lexical bundles in example section

In the presenting content bundles, the lexical bundles with the description function is the most widely used bundles in terms of frequency of occurrence, followed by the location and stance function. Then, it was followed by a group with quantification and procedure functions where they have a small difference in number. Meanwhile, both the framing and the structuring bundles are the least used bundles. In the organizing text bundles, the transition bundles have the highest usage rate. For more details, the nine functions can be seen in Figure 8.

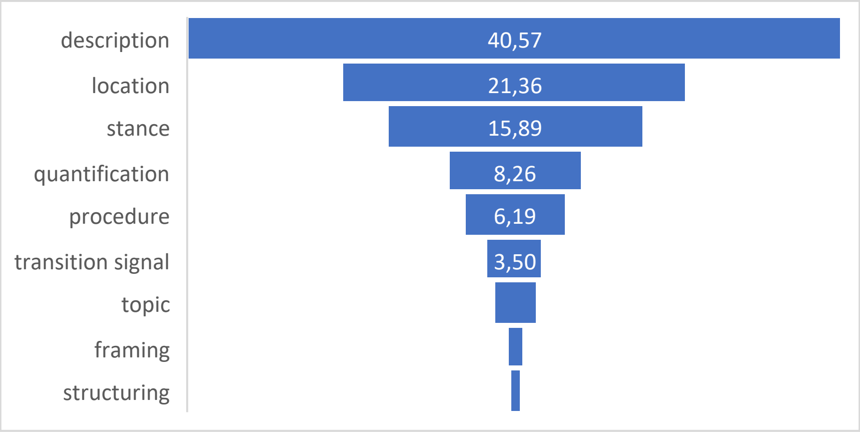


Figure 8 The function of lexical bundles in example section based on occurrence

In terms of variation, lexical bundles with the description function is the most varied bundle, followed by the location and stance function. Thereafter, lexical bundles with quantification, procedures, topics, and transitions functions are in the middle. Meanwhile, structuring function is the lowest level of variation. In the organizing text bundles, the bundles with transition signal has the highest level of variation. The following is the ranking of the nine functions of lexical bundles in terms of variation.

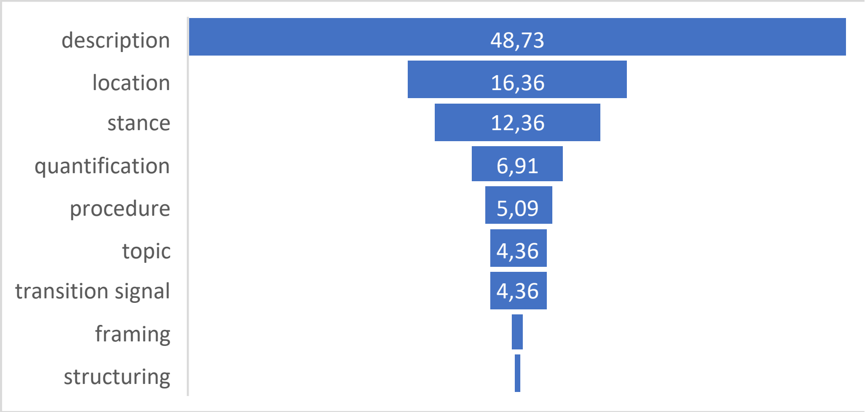


Figure 9 The function of lexical bundles in example section based on variation

4 CONCLUSION

This study aims to find the use of lexical bundles in KBBI Daring, especially in the lemma, definition, and example section. The results showed that the use of lexical bundles in KBBI Daring was mostly found in the definition section, then in the example section, and the last is the lemma section. The bundles found in was generally in the form of phrase rather than clause. In terms of structure, lexical bundles were mostly incomplete structures. The bundles, either in the definition or example section, were mostly in the pattern of *yang*-clause fragment, such as *yang digunakan untuk*, *yang terdiri atas*, *yang terbuat dari*, *yang berasal dari*, and *yang berhubungan dengan*. In terms of function, it seemed that the definition section and the example section have the same tendency that the description and location function have a significant role in building the reader's understanding of a lemma. This study also found a number of potential lexical bundles for KBBI, such as *oleh karena/sebab itu*, *di samping itu*, *dengan kata lain*, *dalam hal ini*, and *di sisi lain*. Therefore, it is suggested to include them as sublemmas and arrange them based on their core elements: *karena*, *sebab*, *samping*, *kata*, *hal*, and *sisi*.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and application* (hlm. 101–122). Oxford: Clarendon Press.

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford guide to practical lexcography*. Oxford: Oxford University Press.

Badan Pengembangan dan Pembinaan Bahasa. (2021). Kamus Besar Bahasa Indonesia (KBBI) Daring. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23): John Benjamins Publishing.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics, 14*(3), 275-311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Budiwiyanto, A. & Suhardijanto, T. (2019). Frequency and structure of Indonesian lexical bundles on academic prose in legal studies: A driven-corpus approach Research Article. *Proceedings of the Third International Seminar on Recent Language, Literature, and Local Culture Studies*, BASA, 20-21 September 2019, Surakarta, Central Java, Indonesia.

Budiwiyanto, A. & Suhardijanto, T. (2020). Indonesian lexical bundles in research articles: Frequency, structure, and function. *Indonesian Journal of Applied Linguistics* 10 (2), 292-303.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*(5), 31-64.

Conrad, S. M., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K. (2009). *Academic discourse: English in a global context*. London; New York: Continuum.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169. doi:10.1017/s0267190512000037.

Jalilifar, A., Ghoreishi, S. M., & Roodband, S. A. E. (2017). Developing an inventory of core lexical bundles in English research articles: a cross-disciplinary corpus-based study. *Journal of World Languages, 3*(3), 184-203.

Kwary, D. A., Ratri, D., & Artha, A. F. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics, 7*(1).

Novita, H. & Kwary, D. A. (2018). Comparing the use of lexical bundles in Indonesian-English translation by students translator and professional translate. Translation & Interpreting Vol 10, No 1 (2018)

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. Amsterdam: John Benjamins.

Samodra, M. C., & Pratiwi, V. D. R. (2018). *Lexical bundles in Indonesian and English undergraduate thesis abstracts.* A paper presented at the 4th PRASASTI International Conference on Recent Linguistics Research.

ENGLISH-JAPANESE DICTIONARIES FOR LEARNERS: PHRASEOLOGICAL PROBLEMS AND PROPOSED SOLUTIONS

Ai Inoue
Tokyo University
inoue125@toyo.jp

Abstract

Referring to phenomena in contemporary English and based on corpus pattern analysis (CPA), this paper aims to demonstrate multi-word expressions' (MWEs', i.e., phraseological units') descriptions in English–Japanese dictionaries for learners (EJDLs). My investigation so far has discovered no published EJDLs perfect in their English phraseology. In fact, they have the following common quantitative and qualitative phraseological problems: (i) the need to review and correct already included MWEs, because some seem old-fashioned, obsolete, and far from correct, and (ii) the need to add newly observed MWEs because finding that a known MWE is neither included nor explained in EJDLs (e.g., *keep somebody in the loop*, *until to* (Inoue 2011), *the way how* (Inoue 2017)) is not surprising. If these two problems are remedied, MWEs in EJDLs would be enhanced both quantitatively and qualitatively. The foremost solution for (i) and (ii) is to reach consensus on a clear definition of MWEs, because EJDLs' definitions of MWEs differ from dictionary to dictionary. Some EJDLs regard a word-combination as an MWE but others do not. The solution for (i) is thus to describe MWEs as they are really used, adopting CPA, although this task would be laborious. The solution for (ii) is to find newly observed MWEs by reading books, articles, etc. and then investigating those MWEs' syntactic and semantic features through CPA. This study begins by defining MWEs and investigating their current situations in well-known EJDLs. Then, the study includes problematic descriptions of MWEs in EJDLs (i.e. (i) and (ii)), along with better and correct MWE descriptions suitable for Japanese English learners. Additionally, the study screens MWEs to find essential and useful ones referring to idioms, collocations, and phrasal verbs in the dictionaries published so far.

Keywords: phraseology, English–Japanese dictionary for learners (EJDLs), multiword expressions (MWEs), Corpus Pattern Analysis (CPA)

1. Introduction

A research field in applied linguistics, second language acquisition, has demonstrated that the reason why individuals employ a language at speed is the use of MWEs (MWEs; i.e. phraseological units), and this leads to obtaining Englishness to learn and use MWEs. How to acquire MWEs and describe them effectively in dictionaries for English learners is a main theme in English lexicography. However, despite the research on defining and classifying MWEs, the theme has remained unsolved. Consequently, this study reviews the history of MWEs' treatments in English–Japanese dictionaries for learners (EJDLs), verifies the treatments of MWEs in recently published EJDLs, and proposes how to provide effective descriptions of MWEs in EJDLs by using data obtained from large-scale corpora and adopting corpus pattern analysis.

This paper is organized as follows. Section 1 overviews the study. Section 2 introduces how English lexicography and phraseology in Japan have evolved. The main topic of Section 3 is to raise problems the EJDLs must solve. Research methods such as defining and categorizing MWEs and the CPA adopted in this study are explained in Section 4. Section 5 introduces data used in this study. Section 6 proposes a solution for providing qualitatively and quantitatively improved descriptions of MWEs. Section 7 sums up the study.

2. English lexicography in Japan

Today's EJDLs are deeply rooted in those published long ago. Yagi (2006: 12ff.) classifies EJDLs into six periods to verify their influences from a historical perspective: (1) a primordial period when collected only vocabularies until 1861, (2) a translational period from 1863 to 1910, (3) a period when English–Japanese dictionaries were published from 1911 to 1926, (4) a period of co-edited English–Japanese dictionaries and a first period when EJDLs were published from 1927 to 1966, (5) a second period when EJDLs were published from 1967 to the present, and (6) a period when unabridged English–Japanese dictionaries were published from 2000 to present. Yagi (*ibid.*) stresses that English–Japanese dictionaries have played a provocative role in understanding English and reaches that conclusion by minutely examining the English–Japanese dictionaries in each period. He also mentions that their substantial appeal is the addition of new vocabularies or new meanings of a word and that it seems that reviewing and correcting descriptions and explanations of words, phrases and so forth in English–Japanese dictionaries leave behind. Hence, he postulates that English–Japanese dictionaries should provide correct information on English.

Similar to the unique history of EJDLs, Japan has its history of English phraseology. Cowie (1999) mentions that H.E. Palmer and A.S. Hornby started English phraseology in the 1920s in Tokyo, but according to Yagi and Inoue (2013: 31ff, 59ff.), English phraseology had been practised in 1909 in a dictionary entitled *A Dictionary of English Phrases*, edited by Naibu Kanda and Tsunetarou Nannichi. The dictionary was compiled based on the educational notion that English phraseological units and basic grammatical rules were essential for increasing the English competence of Japanese English learners. Succeeding the dictionary, *Saito's Idiomological English–Japanese Dictionary*, by Hidesaburo Saito, was published in 1915; *A Second Interim Report on English Collocations*, by H.E. Palmer and A.S. Hornby, was published in 1933; and *Kenkyusha's Dictionary of English Collocations*, by Senkichiro Katsumata, was published in 1939. The dictionaries represent pioneering research in English phraseology and have influenced the development of English phraseology and the treatment of it in English– Japanese dictionaries (Yagi and Inoue 2013). I briefly explain each dictionary in Section 3.

English–Japanese dictionaries and English phraseology in Japan have a long- established history; thus, many imaginative English–Japanese dictionaries have been published. Regarding English phraseology, it converges in the 1990s because of the advancement of computer corpora from the perspectives of educational and linguistic research. EJDLs published after the 2000s have attempted to enrich descriptions of MWEs by using various terms for MWEs; however, the MWEs included are not consistently explained and do not fulfil the need of the times. Overcoming these two limitations would facilitate smooth communication in English.

2.1 English phraseology in Japan

This section introduces how MWEs have been described in three pioneering English phraseological dictionaries in Japan: *A Dictionary of English Phrases* (1909), edited by Naibu Kanda and Tsunetarou Nannichi; *Saito's Idiomological English–Japanese Dictionary* (1915), edited by Hidesaburo Saito; and *Kenkyusha's Dictionary of English Collocations* (1939), edited by Senkichiro Katsumata.

2.1.1 *A Dictionary of English Phrases* (1909), edited by Naibu Kanda and Tsunetarou Nannichi

A Dictionary of English Phrases (*DEP*) is the first phraseological dictionary published in Japan. According to its Preface, the motive to publish *DEP* is as follows:

‘The lack of a reliable phrase dictionary on any comprehensive plan has been regret and a check to thousands of students in the way of learning English’.

Senkichiro Katsumata, discussed in 2.1.3, is not listed as one of *DEP*’s editors, but his methods of selecting and arranging MWEs were adopted in *DEP*. *DEP* presents many idioms used in the sixteenth century, and they are classified into four speech labels according to the editors’ judgements: informal, ill-mannered, obsolete and rare. *DEP* describes the meanings but does not include examples of idioms. Additionally, *DEP* includes 78 registers, and most are not used routinely, such as laws, bookkeeping, geometry, medical science, and rhetoric. Hence, *DEP* is not an idiom dictionary for Japanese English learners but an idiom dictionary for Japanese workers who use English for occupational purposes. *DEP* includes MWEs ranging from technically used idioms related to the registers to collocations such as *abide by* and formulae such as *That’s about it*.

A typical example is *go*. Recently published EJDLs include many idioms under the entry of *go*. Similarly, *DEP* describes various English phraseological units starting from *go a wool-gathering* and *go to bed* from page 707 to page 732, as shown in (1).

(1) ～wool-gathering.途方にくれる、思ひ惑ふ.—To be perplexed.

～aback.退く. ーTo retreat.

～ a bagging.通（ハ）けずにゐる；望み手がない.—To be in no demand; have no applicants: as, last week, strawberries *went a-begging* in the market.

～ aboard. 乗船す；搭乗す.—To enter a ship; embark.

～ about [*prep*.]. ❶…に奔走す、…に取掛かる、…に着手す. ❷[廢]…を求む、

…を追求す.ー❶To busy oneself about, set to work upon, take in hand. She *went about* her work in a cold, impassive way. *Mary Linskill*.

❷To seek after.

Lust is unsatisfiable; to *go about* it is to go about an endless piece of work. *Trapp*.

As (1) shows, an entry, a translation and a definition of a word are sequentially presented. As *go a bagging* shows, an example is described after ‘as’. The examples in *DEP* are not presented with a Japanese translation but with their sources and authors.

Regarding arranging MWEs, *DEP* adopts Katsumata’s method, namely, a noun- oriented principle; thus, an MWE is described in a noun that is one of the components of an MWE. However, the method is not always coherently complied with. For example, collocations comprising a verb and a preposition (i.e. *go to*), of a verb, preposition and noun (i.e. *go to the hammer*) are described in the entry under each verb. In general, *DEP* complies with the noun-oriented principle, clearly arranges the idiom and contributes fulfilling descriptions of idioms. According to the Preface in *DEP*, the editors consulted *Oxford English Dictionary*; *Century Dictionary*; *Brewer’s Dictionary of Phrase and Fable* (first edition in 1870); and *Dictionary of Idiomatic English Phrases*, edited by J.M. Dixon.

2.1.2 *Saito’s Idiomological English–Japanese Dictionary* (1915), edited by Hidesaburo Saito

This section explains why Hidesaburo Saito regarded idiomology (i.e. phraseology) as critical. He wrote about his motive for publishing *Saito’s Idiomological English– Japanese Dictionary* (1915) (*SIEJD*) in its Preface. (Please see (2).)

(2) The following proposition was made at the Second English Teachers' Conference:

> 'That an Ideal Dictionary be compiled. Words are nothing in themselves, and everything in combination. In the case of words, a combination comprises construction and association. A verb without its constructions is no verb (動詞ハ不 動詞); and association is what makes the most significant words what they are. By association are meant the *idiomatic, proverbial* and *conventional* expressions in which each word occurs. The dictionary required is one that shall be the *ne plus ultra* of accurate translations, with the definitions in rational, systematic, genetic order—each word being presented in all its idiomatic, proverbial and conventional associations'.

What is clear from (2) is that idiomology indicates that word combinations including idiomatic, proverbial and conventional expressions and word combinations are critical in English lexicography from an English educational perspective. However, at that time, no persuasive research had been presented on word combinations. Saito mentions that research on word combinations captured his interest and that he was privileged because he could be a pioneer in a novel field of investigation.

Additionally, Saito explains idiomology by reducing it to science, as shown in (3).

(3) Ordinary chemistry has to do with dead matter, and yet their (=combinations) subtlety frequently defines analysis. Our chemistry, on the other hand, deals with living mind, with the action of the human soul, which now shows itself as intellect, now assumes the form of emotion, and now asserts itself as volition.

Thus, idiomology shows not inorganic word combinations but organic word combinations reflecting human beings' feelings. (4) is an easier-to-understand explanation of idiomology than (3) is.

(4) And with all these tasks on his (=a Japanese teacher of English) hands, he must not lose sight of another fact - a wonderful fact with which I ought to have commenced - a fact which is indeed startling in its simplicity, and becomes still more startling when we come to think how some of us seem to ignore it entirely - I mean the fact that *language is made up of words*, which *words* are to the structure of the *language* what the material *elements* are to *chemistry*.

Idiomology and idiomological were coined by Saito. He did not clearly define these terms, but according to the Preface in *Advanced English Lessons* (1901–1902) in (5), they are comparable to today's English phraseology.

(5) It is true that there is English Grammar; but, as it is generally taught and studied, it is nothing more than a set of rules dealing with mere form without matter, and it is justly condemned as being rather a hindrance than a help to the acquirement of the living language. No grammar, rhetoric, or lexicon in existence treats of the living physiology of the language, the multifarious functions of each individual word, the nice distinctions and delicate shades of meaning peculiar to each word and phrases, the spirit and genius of the English idiom. It is not a sufficient explanation to say that an expression is idiomatic. Idiom is a growth, and all growth is subject to natural law. Some idioms have arisen from a tendency to brevity, others from considerations of emphasis and still others from the necessity of distinction. The study of formation of idiom reveals that language, as it is, has not been formed at random, but that the expressions of human thought is governed by laws of economy no less rigid than those which regulate the material world.

Additionally, Saito insists that Japanese English learners should learn idiomology.

The *SIEJD* includes various types of MWEs, as shown in (6). The MWEs described in the *SIEJD* are divided based on the definitions and are explained in Section 3, and (6f) is defined by the author.

(6)	a. (pure) idioms	e.g. *kick the bucket*
	b. proverbs	e.g. *It is no use crying over spilt milk.*
	c. phrasal verbs	e.g. *cry off*

d. collocations e.g. *open an account*

e. formulae e.g. *Day breaks (or dawns)*

f. 熟語 (jyukugo = figurative idioms) e.g. *be in the air*

(6) shows that the *SIEJD* includes all MWEs discussed in present-day English phraseology. Jyukugo (=figurative idioms) in (6f) signifies word combinations in which a keyword among the components is figuratively used. For example, *air* in *be in the air* is used to mean ‘public’. In this case, word combinations are labelled as ‘jyukugo’. Consequently, the *SIEJD* is superior in analysing and describing MWEs, although it was published more than 100 years ago. In addition, MWEs in the *SIEJD* are arranged by degree of idiomaticity[1]. (Please see (7).)

(7) a. idioms high
b. proverbs ↑
c. jyukugo idiomaticity
d. phrasal verbs ↓
e. collocations
f. formulae low

Jyukugo is located in the middle of idiomaticity; thus, it is vague and difficult to understand. That is, idiomology means semantically irregular word combinations. A safe assumption is that Saito considered semantically irregular word combinations (i.e. *jyukugo*) to be the most important type of word combination to learn for Japanese English learners. Saito thought that there were many beneficial word combinations in addition to the semantically irregular word combinations; hence, the *SIEJD* describes many MWEs and has influenced English phraseology in Japan.

2.1.3 *Kenkyusha’s Dictionary of English Collocations* (1939), edited by Senkichiro Katsumata

Kenkyusha’s Dictionary of English Collocations (*KDEC*) was compiled by S. Katsumata and published in 1939. Katsumata stated the characteristics of this dictionary in the Introduction (originally in Japanese): ‘All collocations contained in the dictionary were collected by Katsumata himself and he accumulated collocations from written materials used only by native speakers of English etc.’. The first edition of *KDEC* contained approximately 120,000 collocations. The arrangement of collocations started with a noun to a verb, adjective, or adverb. This method of ordering collocation is called a noun-oriented principle. As aforementioned, this device had been adopted in the *DEP* compiled by N. Kanda and T. Nannichi, published in 1909 and revised by Katsumata. Katsumata published the *Eiwa Katsuyou Gosenku* in 1918. It included 5,000 collocations that mainly comprised transitive verb + object. Using this dictionary as a basis, Katsumata compiled *KDEC*, published in 1939.

The second edition of *KDEC* was published in 1958. The title was changed to *Kenkyusha’s New Dictionary of English Collocations*. The number of collocations increased from 120,000 to 200,000, and increased to 380,000 in the third edition, published in 1995. In addition, 80% of all collocations in the third edition were newly collected. Similar to the first edition, all collocations were based on materials from the writings of native speakers of English. In the noun section, Katsumata focused on collecting the transitive verb + object pattern because he posited that most collocations had this pattern.

Headwords have three clauses: noun, verb and adjective. Each clause has subcategories. For instance, nouns have V, V^2, Q and Q^2. From the second edition, the code Q^2 was introduced to indicate noun +

noun collocations. In addition, verbs include M and M^2, and adjectives are coded by M and P. These subcategories are code patterns, and all collocations have code patterns. Code patterns employed in this dictionary are more intricate than those of *The BBI Combinatory Dictionary of English* (*BBI*). The definition changes for each example. This dictionary explains which words each headword collocates with. These characteristics substantially differ from those of *BBI*.

Compared with *BBI*, *Kenkyusha's New Dictionary of English Collocations* has much more detailed information for each headword. In the third edition, many collocations and examples were added. Moreover, each collocation was checked and revised by native speakers of English, making the dictionary more satisfactory in quality and quantity than other collocational dictionaries. Katsumata claimed that this dictionary was a pioneer among collocational dictionaries and was second to none compared with foreign collocational dictionaries. The only faulty aspect was its scarcity of spoken collocations.

3. Problems with EJDLs

Based on my review of recently published EJDLs, to overcome their limitations, I propose three solutions: EJDLs should qualitatively and quantitatively fulfil the descriptions of MWEs, review the MWEs described in EJDLs, and because some are old-fashioned, add new MWEs.

To realize the first solution, EJDLs should adequately define MWEs and their subcategories and then select and describe the MWEs necessary for English learners by referring to MWEs described in idiom dictionaries, collocation dictionaries, phrasal verbs' dictionaries and other phrase-related dictionaries. To realize the second solution, EJDLs should find their incorrect or old-fashioned MWEs. This topic might be an interesting lexicographical and phraseological topic for investigating what triggers such incorrect or old-fashioned MWEs, but this topic is beyond the scope of this study. Lexicographers may need time to provide accurate descriptions of MWEs by adopting the method of CPA. To realise the third solution, EJDLs should find newly observed word combinations by reading a magazine or newspaper and judging whether a word combination is an MWE by using the four criteria explained in Inoue (2018b) and then adding the newly observed MWEs to EJDLs. The third solution demonstrates that the qualifications of a lexicographer are tested. I will discuss the second and third solutions, including concrete examples, and present appropriate descriptions of MWEs in EJDLs.

3.1 MWEs–their definitions and problems to be fixed

No clear, non-controversial definition of English phraseology has been provided. A vague definition would be that English phraseology is the study of phrases. However, this definition is unsatisfactory because which phrases are included or excluded is not described. Hence, this study defines English phraseology as the study of repeatedly used phrases comprising at least two words. The definition includes word combinations such as idioms, collocations, phrasal verbs, proverbs (=sayings), formulae, discourse particles and fixed phrases. The umbrella term for such word combinations is phraseological units (PUs). Phraseme is another term for PUs. The two terms vary by the study because the targeted word combinations differ. This study refers to PUs as MWEs and defines each word combination based on the following standards: the frequency, polysemy, semantic transparency and commonly used definitions of each word combination. Notably, this study does not intend to fully explain the definitions of each word combination by comparing them to those in the literature.

Idioms such as *keep one's head* are not frequently used. Their meanings are not the sum of each component; thus, idioms are not polysemous word combinations.

Collocations (e.g. *set up*/*launch a company*) range from high to low in frequency. In both cases, they are not polysemous and are semantically easily predictable from each component. For example, *set up a company* is more often used than *launch a company*.

Phrasal verbs are word combinations comprising either a verb and an adverb or a verb, an adverb (optional) and a preposition. Phrasal verbs are frequently used but are not polysemous word combinations. Semantically, phrasal verbs are not always composed by the sum of each component. For example, *look around*, *look up to* and *put off* are phrasal verbs.

Formulae, for example, *now you're talking*, *Thank God/Goddess, it's Friday*. and *I wasn't born yesterday*. appear in a conversation and do not have polysemy. Their frequencies differ by formula. Additionally, some formulae are semantically easy to understand, but others are not.

Proverbs (=sayings) such as *Don't teach your grandmother to suck eggs* are not the sum of each component from a semantic perspective. Proverbs are used in a limited context; thus, they are neither frequently used nor polysemous word combinations.

Discourse particles have a high frequency when used in a conversation and have polysemy. For example, discourse particles such as *you know*, *I mean* and *let's see* have both a literal meaning and a pragmatic meaning in accordance with the context in which they are used. Discourse particles, such as after all and and stuff like that, are semantically difficult to understand because they are not the sum of each component and are not polysemous word combinations.

Last, Inoue (2007) discusses fixed phrases, which have high frequency and polysemy, such as *you know what*, *here we go (again)* and *let's say*. Some fixed phrases such as *until before* and *until by* discussed in Inoue (2019) are monosemous (i.e. antonym of polysemous). A common aspect of monosemous and polysemous fixed phrases is that they have been overlooked in the literature because they are formed beyond the explanations of theories and English grammatical rules.

The aforementioned MWEs can be classified into two categories, as presented in (8): MWEs that can be explained and are within the theories and English grammatical rules (i.e. the former six word combinations are in this under the category), and MWEs beyond the explanations of theories and English grammatical rules and referred to as irregularities, of which only fixed phrases are deemed.

(8) a. word combinations not beyond the explanations of the theories and English grammatical rules (i.e. idioms, collocations, phrasal verbs, formulae, proverbs and discourse particles)

b. word combinations beyond the explanations of theories and English grammatical rules (i.e. fixed phrases)

4. Research method

The study adopts CPA, a procedure used in corpus linguistics that associates word meaning with word use by analysing phraseological patterns and collocations.

According to Pustejovsky *et al*. (2004), in CPA, the meaning of a pattern is expressed as a set of basic implicatures. For example, for the verb *file*, one pattern is [[Human = Plaintiff]] file [[Procedure = Lawsuit]], for which the implicature may be expressed as *If you file a law suit, you are acting as the plaintiff and you activate a procedure by which you hope to obtain redress for some wrong that you believe has been done to you*. Depending on the proposed application, the implicature of a pattern may be expressed in various other ways, for example, as a translation into another language or as a synonym set such as 'file = activate, start, begin, lodge'.

5. Source materials: EJDLs and data used in the study

This article deals with phraseological problematic descriptions in four major ELDLs, which were recently published. The availability of large-sized computer corpora has made it possible to provide insights into phraseological research. This study uses the data from the Corpus of Contemporary American English (COCA), British National Corpus (BNC), WordBanks*Online* (WB) I accessed COCA on the 12th, 13th and 14th of February, 2021, BNC and WB.

6. A solution proposal–how to effectively describe MWEs in EJDLs

Unsurprisingly, in EJDLs, I occasionally observe MWEs that do not adhere to their actual behaviours in contemporary English. This section proposes a solution to the problem of how to correctly incorporate MWEs, which seem essential to English learners, into EJDLs.

6.1 Old-fashioned MWEs: idioms and collocations

One of the MWEs in EJDLs, idioms and collocations is qualitatively and quantitatively relatively fulfilling and rich, but this situation backfires on EJDLs in that they include some old-fashioned MWEs. For example, almost all EJDLs still describe the well-known idiom *rain cats and dogs* (e.g. *It's raining cats and dogs*.) by using labels *old-fashioned* or *informal* under the entry of *rain.* The description might confuse Japanese English learners regarding how they express that it rains heavily with Englishness. Advanced English learners can easily come up with *It's pouring down*, but 80% of Japanese English learners belonging to Level A of CEFR (Common European Framework of References for Languages: Learning, teaching, assessment) offer little hope that they use MWEs with Englishness. EJDLs separately describe *It pours/is pouring down* under the entry of *pour* from *rain cats and dogs* and do not explain that *It pours/is pouring down* is more common than *It's raining cats and dogs*. To vary the description in EJDLs, as (9) shows, it is user-friendly to place *It/The rain pours/is pouring down* and *It's raining cats and dogs* together under the entry of *rain* with appropriate labels.

(9) It's/The rain is *pour*ing *down*. (= *It's raining cats and dogs*. (old-fashioned, informal))

Users of EJDLs are mainly digital natives, and EJDLs include unfamiliar collocations for them, such as *rewind a video*. Such MWEs seem difficult to understand and imagine for digital natives. When *rewind a video* changes into *rewind a film*, it helps users easily understand what is occurring.

Such old-fashioned MWEs are too numerous to mention in EJDLs; thus, EJDLs are not always user-friendly, because they do not cross-reference MWEs such as *rain cats and dogs* and *pour down*. One of the urgent agendas of EJDLs is to vary old-fashioned MWEs' descriptions into those which match the times.

6.2 Incorrect MWEs

EJDLs sometimes explain the different meanings and functions of MWEs from those actually used in a context in contemporary English. For instance, EJDLs describe *train of thought* (e.g. *I've lost my train of thought*.), but data from corpora show that *be in the same train of thought* is a commonly used pattern.

6.2.1 Semantically similar MWEs

EJDLs separately describe *here we go* and *here we go again*, but Inoue (2007) and Yagi and Inoue (2013) have demonstrated that some functions of the two MWEs overlap (Table 1).

Table 1 Polysemy of *here we go* and *here we go again*, and their syntactic and phonetic characteristics

	typically co-occurring words and phrases	***position in the sentences***	***tone***	***form***
to call for attention	look, listen	beginning		here we go
to rouse people to do something	OK, all right are you ready?	beginning falling	rise-fall here we again	here we go, go
to express irritation	oh, no	beginning	here we again	here we go, go
to express agreement		middle	pause before here we go	here we go
to find something	OK	middle		here we go
to show something		middle		here we go

(Inoue 2007: 169)

On the other hand, EJDLs regard the idiom *learn the ropes* same as *know the ropes*, but Inoue (2021a) reveals from data obtained from corpora that *know the ropes* in (10) does not behave same as *learn the ropes* and that *find the ropes* in (11) is a derivate from *learn the ropes*.

Know the ropes is used to say to 'be thoroughly versed in something and acquire it in the light of various experiences'; hence, it semantically differs from *learn the ropes*. *Know* in *know the ropes* is used to mean 'be familiar with something including advantages and disadvantages and master it' and shows that the cognitive status takes a step forward from *learn*. In addition, *ropes* in *know the ropes* signifies not tips but the entire picture of something including merits and demerits. In other words, *ropes* in *know the ropes* semantically extends from *ropes* in *learn the ropes*. Consequently, *know the ropes* causes the semantic extension of *learn the ropes* and cognitively moves to a subsequent phase (i.e. be familiar with and acquire something including pros and cons). In addition, *learn the ropes* is not replaced with *know the ropes*, unlike dictionaries' descriptions.

(10) a. You're a respected senior senator, and you *know the ropes*, you know your job extremely well. (COCA, 2019, FIC)

b. I am genuinely pissed off. I am a liberal democrat who was shaken to the core by 9/11. I was ready to back the administration in pursuit of those responsible. With 96 combat missions, two space flights, and retired CEO of a Defence Department think tank, I *know the ropes* and the risks. (COCA, 2012, WEB)

(11) a. Being a criminal was preferable to being a deadbeat like his own father, whom he last saw when he was around 5, he said. (Offset's first felony conviction, in 2012, was for possessing stolen property.) However he also maintained that his home life was not to blame for his wayward years. 'That was me being a knucklehead, trying to *find the ropes*,' he said. (COCA, 2018, NEWS)

b. The damage, though, was inflicted between the seventh and tenth overs which yielded only 12 runs against the spin of:PERSON: and:PERSON2: PERSON:, below, finished at better than a run a ball but struggled initially to *find the ropes* whether through lack of timing or a lack of power. (WB, 2014, TIMES)

Find the ropes means 'acquire tips in an effort' because expressions such as *trying to* and *struggled* typically co-occur with *find the ropes*. In other words, *find the ropes* is used to say that a speaker voluntarily attempts to acquire tips with an effort but is not sure if s/he can obtain tips in the end and whether s/he can help to do better something related to a job. In summary, *find the ropes* is derived from *learn the ropes*, but it is unpredictable that *find the ropes* eventually reaches *know the ropes*.

6.2.1 Lack of Englishness

Some Japanese English learners misuse *I'm jealous* and *I envy you* because they seem not to know the difference between *jealous* and *envy*; thus, their utterances lack Englishness. A similar situation is shown in (12). I asked Japanese college students to choose the correct answer to (12a).

(12) a: How's it going?

b: I got a problem./I got trouble.

Some of the students choose the right answer *I got a problem*, and others choose *I got trouble*. I asked all the students why *I got a problem* was the correct answer, and they could neither provide an explanation nor understand the differences between *problem* and *trouble*. *Problem* is used to mean a question to be considered, solved, or answered, but *trouble* is used to say a state of distress, affliction, difficulty, or need. If English learners do not know the difference, their response to (12a) sounds awkward. Additionally, they need to learn that *trouble* is used as in *get into trouble*.

Next, I present an example of an MWE for which a wrong construction is used, leading to a lack of Englishness. When we use the MWE *tear one's hair (out)*, Japanese English learners tend to use the MWE in the construction, *I'm tearing my hair (out) over the problem*, but *The problem has me tearing my hair (out)* is more natural. EJDLs should explain which construction is compatible with MWEs in examples.

Inoue (2021b) demonstrates that Japanese English learners cause discrepancies in input-output English and Japanese MWEs. The cause of the discrepancies seems to be that the input-output MWEs are unconsciously influenced by the usages of the MWEs used in the first language and by the cultures of the first language.

First, I asked Japanese postgraduate students who have learned English phraseology for one year and are English learners to translate the idioms shown in (13) into Japanese without explaining the meanings of the idioms in (13) (italicised by the author.). Almost all the L2 learners could not correctly translate the idioms by sight.

(13) a. The actor *shook his head* at the offer to play a leading role in a stage.

b. The famous actress was arrested for the use of cannabis. It was very difficult to*wash her hands of* cannabis once she started using it.

c. The new personnel *has* such *a loose tongue* that we cannot tell him our crucial matters.

d. She does *not bat an eye* no matter what happens to her. She is really something.

e. She has been working here for more than ten years, so she'll *show the newcomer the ropes*.

f. All employees' *jaws drop to the floor* because of his outrageous words and actions.

g. It is said that this charm is *a rabbit's foot*.

h. Government employees should conduct themselves *by the book*.

Next, I explain to the L2 learners the idioms used in each sentence, for example, their meanings, origins and syntactic features as follows: generally, *shake one's head* is translated into 首を横に振る (*kubi wo yoko ni furu* means shake crossly one's neck, *kubi* means neck) in (13a), but in English *head* instead of *neck* is used, for example, *shake one's head*. As for (13b), *wash one's hands of* is translated into 足を洗う

(*ashi wo arau*, *ashi* means legs or feet). Different body parts, *hands*, are used in English. The idiom *have a loose tongue* in (13c) is equivalent to 口が軽い (*kuchi ga karui*, *kuchi* means a mouth) in Japanese. Same as (13a, b), a different body part is used in English. *Not bat an eye/turn a hair* in (13d) is the same as the idiom 顔色を変えない (*kawo iro wo kaenai*, *kao* means a face). From (13a) to (13d), different body parts from English idioms are used in Japanese idioms. *Show someone the ropes* in (13e) is derived from seafaring things, and *the ropes* is figuratively used to mean a tip. *Teach someone the ropes* and *know/learn someone the ropes*, as included in EJDLs, can be used in a workplace instead of *show someone the ropes*. The idiom *jaws drop to the floor* in (13f) is used by Justin Pierre James Trudeau, Canadian Prime Minister, to express astonishment at former U.S. President Donald Trump talking with other countries' prime ministers and presidents like *His teams' jaws dropped to the floor*. The idiom *jaws drop to the floor* is used for saying that someone is very surprised and shocked, and it is not included in major ELDLs. EJDLs merely describe that *jaw* is used to show surprise or disappointment. In the case of (13g), *a rabbit foot* is not included in EJDLs, although it has been said to bring someone good luck in English. *By the book* in (13h) is explained as correctly following rules or systems for doing something in a strict manner in EJDLs and is substituted as *according to the book*.

Last, a couple of weeks later, without advance notice, I asked the L2 learners again to translate the idioms in (13) into Japanese and to explain the idioms. Next, I analyse and investigate how the expressions are unconsciously influenced by a native language from linguistic and cultural standpoints. Table 2 shows the expressions translated by the postgraduate students.

Table 2 Translation of familiar English idioms into Japanese by L2 learners

	equivalent and correct Japanese idioms to the English idioms in (13)	idioms translated by L2 learners
(13a)	首を横に振る	頭を振る(*atama wo furu*, 頭(*atama*)=head) is the literal meaning of *shake oneś head*.
(13b)	足を洗う	手を洗う(*te wo arau*, 手(*te*)=hand) is the literal meaning of *wash oneś hands of*.
(13c)	口が軽い	no answer
(13d)	顔色一つ変えない	瞬きしない(*mabataki shinai*, *mabataki* = blink) is the literal meaning of *not bat an eye*.
(13e)	コツを教える	コツを教える
(13f)	あんぐり口を開ける	あごが外れる(*ago ga hazureru*) means dislocate one's jaws.
(13g)	幸運をもたらす	no answer
(13h)	規則に従って	本に書いてあるように(*hon ni kaite aruyouni*, 本(*hon*)=the book） is as the book writes.

The results reveal that L2 learners could not correctly translate the English idioms into Japanese, although I had explained them a couple of weeks ago. In (13a, b), the L2 learners literally translate the two English idioms into Japanese; thus, they wrote body parts that differed from those in the original English idioms. In (13c), L2 learners did not recall any equivalent Japanese idioms to *have a loose tongue*; thus, they could not translate it into Japanese. As for (13d), L2 learners posited blink from the component *an eye* in *not bat an eye*; thus, the translated Japanese is not idiomatic. In (13e), L2 learners correctly understood the idioms, and some L2 learners wrote the alternative verbs of *show* in (13e). In the case of (13f), L2 learners wrongly translated *jaws drop to the floor* into あごが外れる (lit. dislocate one's jaws)*; thus,* it did not make sense. An L2 learner translated the idiom (13f) into 笑いすぎてあごが外れる(lit. dislocate one's jaws because of too much laughing). *Dislocate oneś jaws*（あごが外れる）has two meanings in Japanese: one meaning is literal, and the other meaning is idiomatic and used to express that something is so funny for somebody that s/he almost has a dislocated jaw because of too much laughing. The reason why L2 learners did not correctly translate the English idioms is that Japanese does not have semantically equivalent idioms to (13f). Additionally, in (13f), *mouth* is used instead of *jaws* in Japanese (e.g. あんぐり口を開ける, lit. a mouth wide open with surprise, astonishment, etc.); thus, it is difficult for Japanese to deduce the meaning of *jaws drop to the floor* from *jaws*. In (13g), L2 learners could not understand and translate the idiom because, as I have explained, because of Japanese folk stories, *rabbits* have been long considered evil. Hence, this leads to no answer for the meaning of the idiom *a rabbitṡ foot*. As for (13h), L2 learners understood that *book* is used for saying *a rul*e; thus, they could correctly translate (13h) into Japanese. Consequently, Japanese ways of thinking and Japanese cultures subconsciously influence the translation of English idioms into Japanese.

Whatever the case, code-switching from Japanese to English does not properly work, because of the influences of Japanese idioms, ways of thinking, and cultural backgrounds in the case of translating Japanese idioms. Hence, the idioms translated by L2 learners lack Englishness. In addition, EJDLs fail to correctly describe the correspondence relation between Japanese idioms and English idioms, which might lead to insufficient Englishness and losses in translation.

6.3 Newly observed MWEs

A famous MWE *Itṡ (ainṫ) over until the fat lady sings* should be changed to *Itṡ (ainṫ) over the lady sings* because the expression *the fat lady* in relation to operas is inaccurate because sopranos are typically slender and referring to an individual by their weight has become politically incorrect. Similar with the MWE, it is not surprising to find newly observed MWEs in contemporary English, and investigating their actual manners based on the date from corpora is not difficult. I have presented newly observed MWEs, for example, *until to* in (14), *it looks that*-clause in (15), *in accordance to* in (16), *they who* ~ in (17), and *take care for* in (18).

(14) a. This means that average household size in Great Britain fell from about 3.21 to about 2.56 persons over this period and this decline is expected to continue at least *until to* the end of the century. (Inoue 2018b:35; BNC)

b. Follow Hill-Brady Road *until to* the stop light at Dickman Road/M-96 ('Maps and driving directions, Kellogg Hotel & Conference Center Michigan State University; ibid.)

(15) a. With a 34–0 lead, top gear was no longer required. Yet, despite their form over the last two months, the turning point for Wigan's 20th major trophy in six years undoubtedly came in December when *it looked that* they might struggle for success this season. (Inoue 2018b: 90; BNC)

b. 'A girl in my class told me about it,' adds Dolores. 'She knew the lads and said they were very nice so that made it easier to audition for them'. And did the lads turn out to be very nice? She pauses for a moment and the van overflows with the sound of helpless male laughter. 'Well they were townies you know,' she finally says, 'and *it looked* to me *that* when townies hung out together they all dressed the same, did the same things, went to the same places….' (Inoue 2018b: 90; BNC)

(16) a. All animals were cared for *in accordance to* the Guide for the Care and Use of Laboratory Animals as published by the National Institutes of Health. (Inoue 2020a: 2; COCA, 2015, ACAD)

b. They're expected to know how to act *according with* the ethics of the environment. (Inoue 2020a: 2; COCA, 2014, ACAD)

(17) a. Gluttony, inebriety, anger, peevishness, and melancholy, are strong provocatives of the disease,

and *they who* indulge in them may do it at the expense of their lives'. (Inoue 2020b: 33; Laws 2014)

b. MAN: The first gun that was fired at Fort Sumter sounded a death knell of slavery. *They who* fired it were the greatest practical abolitionists this nation has produced. (Inoue 2020b: 33f.; COCA, spoken, 2011)

(18) a. 'You're supposed to select someone who will be good for the baby,' Amy said. 'Someone to look out for and *take care for* the baby. ….' (Inoue 2018a: 2; COCA, 2014, FIC)

b. During instruction, such values and related attitudes can be obtained if several conditions are established: building a community with members who *take care about* each other, …. (Inoue 2018a: 2; COCA, 2005, ACAD)

c. Tebow said. 'And it would just be me and my mom at the house. So, it was my responsibility until I was old enough to go (at age 15) to *care of* the cows, ….' (Inoue 2018a: 2; COCA, 2012, NW)

EJDLs neither describe nor explain newly observed MWEs; thus, research on such MWEs should be continually conducted, and new findings should be correctly described in EJDLs.

6.4 Ideal descriptions of MWEs in EJDLs

EJDLs allocate many pages for explaining essential MWEs for Japanese English learners. These pages follow the special pages in the middle of English dictionaries for learners. The reason is that EJDLs do not cross-reference between semantically similar MWEs, do not explain how subconscious knowledge of Japanese influences the input and output of MWEs used in English, and do not catch up with newly observed MWEs because of the limited space under the entry of a long word.

7. Concluding remarks

English lexicography in Japan has paid attention to MWEs for more than one hundred years through trial and error, but the treatment of MWEs in EJDLs is not always satisfactory and should be revised from quantitative and qualitative aspects. Hence, this article has presented solutions for phraseological problems, namely, provide clear definitions of MWEs and describe MWEs as they are authentically used in a context adopting CPA.

MWEs play a significant role in second language acquisition. In English lexicography, rich, fulfilling MWEs in EJDLs would lead Japanese English learners to communicate while using Englishness. Additionally, this would help Japanese English learners obtain a more advanced level of CEFR. The descriptions of MWEs in EJDLs should be minutely investigated by adopting CPA to make this educational implication more fruitful, although it is laborious for phraseologists and lexicographers to do so. It is high on the lexicographical agenda that EJDLs beings by selecting essential, useful MWEs for Japanese English learners and describing the actual and correct manners of MWEs.

Acknowledgement

This research was supported by JSPS KAKENHI Grant Number 20K00674 (Grant-in- Aid for Scientific Research (C)).

Notes

[1] According to Moon (1998), idioms can be classified into high or low idiomaticity on the basis of three features: institutionalisation, lexicogrammatical fixedness and (semantic) non-compositionality. For example, idioms such as *kick the bucket*, *call the shots* and *kith and kin* have high idiomaticity because they are conventionally and fixedly used, and it is difficult to infer the meanings from each component. By contrast, idioms such as *enough is enough* and *because of* are regarded as having low idiomaticity because the meanings are easy to understand despite being also conventionally and fixedly used. Idioms can be

classified into four types by idiomaticity: free combinations (e.g. *open a window*), restricted collocations (e.g. *meet the demand*), figurative idioms (e.g. *call the shots*) and pure idioms (e.g. *spill the beans*) (Cowie 1999: 71). For additional details, please see Cowie (*ibid*.).

References

Corpora

BNC: British National Corpus (http://scnweb.jkn21.com/BNC2/)

COCA: The Corpus of Contemporary American English (http://corpus.byu.edu/coca/) WB: WordBanks*Online* (http://scnweb.jkn21.com/WBO2/)

Books and papers

Cowie, A. P. (1999). *English dictionaries for foreign learners: A history.* Oxford: Clarendon Press.

Inoue, A. (2007). *Present-day spoken English: A phraseological approach*. Tokyo: Kaitakusha.

Inoue, A. (2011). A phraseological approach to finding the functions of newly observed compound prepositional phrases *until to* and *up until to* in contemporary English. *Lexicography: Theoretical and practical perspective* (ASIALEX'11 Kyoto proceedings), pp. 160–169.

Inoue, A. (2017). Newly observed phraseological units beyond the explanations of existing linguistic frameworks–*the way how* as an example. *International Journal of English Language and Linguistics Research, 5*(3), pp. 1–19.

Inoue, A. (2018a). 'Newly established idioms through the blending of semantically similar idioms - *take care for*, *take care about*, and *care of*.' *Lexicon*, 48, 1–24.

Inoue, A. (2018b). *Eigo Teikeihyougen no Taikeika wo Mezashite–Keitairon, Imiron, Onkyo Onseigaku no Kantenkara* (*Working toward the Systematization of English Phraseology from the Three Perspectives of Morphology, Semantics, and Acoustic Phonetics*). Tokyo: Kenkyusha.

Inoue, A. (2019). 'English phraseological research on *until by*/*before* working as complex prepositions.' *International Journal of English Linguistics*, 9(1), 1–14.

Inoue, A. (2020a). 'A lexical priming's analysis of semantically similar group prepositions in formal English.' *Lexicon*, No. 50, pp. 1–23.

Inoue, A. (2020b). 'Shinriteki kyori to chuushodo ni yoru daimeishi no tsukaiwake ga oyobosu eikyou–hito wo arawasu *they who*, *these who*, *those who* no baai.' (The decision-making of selecting pronouns due to psycological distance and level of abstraction–in the case of *they who*, *these who*, *those who* implying people). *Collected papers from Kansai Linguistic Society for English Grammar and Usages*. Tokyo: Kaitakusha.

Inoue, A. (2021a). 'Are idioms arbitrarily changing? A large-scale corpora investigation' Presented at The International Society for the Linguistics of English (6/2/2021)

Inoue, A. (2021b in press) 'Aspects of multiword expressions in Asian lexicography: Japanese.' Jackson, H. (ed.) *The Bloomsbury Companion to Lexicography* (2nd ed). U.S.: Bloomsbury.

Moon, R. (1998). *Fixed expressions and idioms in English*. Oxford: Clarendon Press. Pustejovsky, J., P. Hanks., and A. Rumshisky. (2004). "Automated induction of sense in context." *COLING 2004.* Switzerland: Geneva.

Yagi, K. (2006). *Eiwa Jiten no Kenkyu–Eigo Ninshiki Kaizen no Tameni* (*Research on English-Japanese Dictionaries–To Improve the Awareness of English*). Tokyo: Kaitakusha.

Yagi, K. and A. Inoue. (2013). *Eigo Teikeihougen Kenkyu–Rekishi・Houhou・Jissen* (*Research on English phraseology–its History, Method, and Practice*). Tokyo: Kaitakusha.

A STUDY INTO ENGLISH WORD-FORMATION COMPARISON BASED ON TWO ENGLISH NEOLOGISM DICTIONARIES

Aling Shi
Fudan University, Shanghai, China
shialing2011@163.com

Abstract:

The rapid change in social life and computer application provide fertile soil for neologism generation. And in turn neologisms enrich vocabularies, promote communication and also draw attention on lexicology study. There are many papers on word-formation of English neologisms but few on the comparative study of word formation based on two English neologism dictionaries. Grounded on the neologisms in *A Supplement to the English-Chinese Dictionary (Unabridged)* and *The English- Chinese Dictionary of New Words*, this paper analyzes the word-formation features in a bid to reveal the similarities and differences between those two dictionaries and tries to dig out the reasons for that. Thus, the word-formation tendency of English neologism is likely to be predicted.

Keywords: English neologism; Word-formation; Comparison

1. Introduction

Language records social changes and its development in a bid to reflect aspects of social life. Everything experiences dynamic variations as new things come into being and old things die out. So does the development of language. The English language is notoriously fast in adapting to the changing world. Every year around a thousand new words enter English from every area of life where they represent and describe the changes and developments that take place from day to day. These new additions show how much the English language has changed over time. Thomas Steams Eliot, English famous modernist poet and literary critic, once remarked that for last year's words belong to last year's language, and next year's words await another voice. Language is in step with social development. "A great number of novel words have entered the English vocabulary since the last century, especially in the 1980s–1990s, with the advent of new technologies and new media, such as the Internet." (Mattiello, Elisa, 2017:26) Compared with pronunciation and grammar, vocabulary changes are easy to be noticed. "The emergence of abundant new words is natural result of the great changes in social economy and culture. For one thing, they manifest the changes in people's lives and thoughts. On the other hand, as the carrier of culture, these new words and expressions leave historical traces for the social development." (Song&Yang, 2006)

The emergence of new words not only enriches vocabulary and promotes communication, but also attracts researchers to study vocabulary. Lexicographical researchers have never stopped studying English neologisms, and have achieved great achievements in research contents with high levels. From the perspective of English neologism and word-formation, previous scholars mainly focused on the formation of English neologisms (Guo Nianzhong 1990), the exploration of the formation of English neologisms based on metaphor (Yu Jing 2007), the adoption of the principles and strategies of English neologism formation and translation (Wang Xiaohan 2014), and the formation trend of English neologisms (Wang

Shuang 2014), the cognition of English neologisms in word formation from the perspective of concept integration (Li Lu 2018), and the comparative study of English and Chinese neologisms in word formation (Hao Yue 2018). However, few scholars took two English neologisms dictionaries texts as research objects to carry out comparatively diachronic studies on English neologism word-formation.

The research objects of this paper are *A Supplement to the English-Chinese Dictionary (Unabridged)* (short for SECD) edited by Lu Gusun and *The English- Chinese Dictionary of New Words* (short for ECDNW) edited by Gao Yongwei. SECD is a dictionary complied by Lu Gusun and his team to supplement the first edition of *The English-Chinese Dictionary*, which includes 3500 new words, new usages and new meanings in all, and 8000 examples that manifest the meaning, usage and social culture embedded in words. The words in SECD mainly is collected from the 1980s to the late 1990s. As for ECDNW, it contains about 4100 new words, new usages and new meanings. "The selection of words is well-grounded, wide-ranging and selective, rather than excessive, not only from the printed text, but also from the common words of Internet and mobile communication, and the documentary evidence is indicated." (Lu Gusun, 2018) The words in ECDNW mainly is collected from the beginning of the 21st century to the end of the 21st century. In this sense, ECDNW is a continuation of SECD. Based on the new words collected in SECD and ECDNW, this paper analyzes the formation of new words and compares the similarities and differences between them in order to understand the rules of English new word formation, thus get to understand English word-formation and English vocabulary as a whole.

2. Word-formation of New Words

English new words are not only increasing in number, but also coming from a variety of sources with concise trend. "Interestingly, it is actually very rare for an English word that is completely new to be formed. Often, repurposing takes place; in other words, a new sense is added to an already existing word." (O'Dell, Felicity. 2016:95) He mentioned about affixation, repurposing, compounding, blending, importing words, and abbreviations as ways to form new words. Marchand Hans (1969) described compounding, prefixation, suffixation, derivation by zero-morpheme, back-derivation, clipping, and blending and word-manufacture in his book *The Categories and Types of Present-Day English Word-Formation.* Cannon, G. (1979) extracted 6000 New English words from *the OED Supplements* and *The Barnhart Dictionary of New English Since 1963*, and found that affixation accounted for a large proportion of new words and compounds. Bauer Laurie (1983) discussed affixation, derivation without affixation, and compounding. There are nine types of neologism based on Newmark's theory. They are old word with new sense, new coinages, derived word, abbreviation, collocation, eponym, acronym pseudo and blends. (Cao, 1984) Adams Valerie (2001) in his book named *Complex Words in English* mentioned word-formation processes in English, including transposition, prefixation, suffixation, formations with particles, and compounds. *In A View into Retronymy as A Source of Neology*, George. J. Xydopoulos and Irene Lazana (2014) specifically discussed word formation related to retronymy. There are also many discussions on the formation of English Neologisms in China. A and Song (1957) first discussed the formation of English compound words. Guoqiang Lu listed six kinds of new words, such as blending, coinage, acronym, back-formation, conversion and loanwords (1996: preface, 1-3); Yonglin Yang (1997) mentioned ten manners to form new words: coinage, derivation, compounding, blending, shortening, conversion, back-formation, antonomasia, initial letters, loanwords; Yihua Zhang (2003) classified the construction methods and word formation motivation of English new words and thought that there were five main word formation methods: compounding, abbreviation (including acronym, BBS is the abbreviation of bulletin board system; acronym, call refers to computer assisted language learning, computer-aided language learning; shortening, ad = advertisement, chute = parachute, scrip = scripture), simile (surfing, from water surfing game to Internet surfing), analogy (such as from user friendly to listener friendly/reader friendly), loan words. Wang (2000) took *The Oxford Dictionary of New Words* as a study objective material, and extracted 2000 English new words, finding that in addition to those newly coined English words such as dweeb and Prozac, most new words are formed in the traditional method which includes the following six categories: compounding, abbreviations, derivation, function shifting, back formation and onomatopoeia. Gao (2001) used the associative cognition of new vocabulary to divide the composition of vocabulary into affixation, compounding transplantation , direct transplantation, blending and acronym. From the perspective of improving reading level, Liu and Ji (2006) divided the word-formation methods of English new words into compounding, derivation and conversion, so as to guide learners to use word-formation method to have a wide guess of new word meanings, and to know the topic of articles and improve reading speed. Cheng and others (2015) took the advantage of the Internet to collect English new words that appeared in the second half of 2013, and screened out 171 new words with distinct parts of speech and definition. By classifying those news words in line with morphology, they found that there were three major ways of new words formation, including affixation, compounding and blending.

Based on the collection of SECD and ECDNW as well as afore-mentioned discussion about word-formation, this paper includes various word-formation methods, new meanings, new usages of old words and loanwords. Among them, various word formation methods mainly include compounding, derivation, conversion, abbreviation, blending and other word-formation ways such as analogy, onomatopoeia, de-nominalization, variation, misspelling, respelling and coinage.

3. Analysis and Discussions of Neologism Word-formation in SECD and ECDNW

The subject of this paper is neologisms in SECD and ECDNW, with their sources of new meanings, new usages of old words and loanwords. New words from each source are shown in Table 1 below.

Table 1 Source Comparison between SECD and ECDNW

Methods for neologisms	SECD	Proportion in SECD	ECDNW	Proportion in ECDNW
Word- formation	2515	87.0%	3795	91.8%
New meanings	164	5.6%	101	2.4%
New usages	72	2.4%	47	1.1%
loanwords	99	3.4%	195	4.7%

The total number of words collected in SECD and ECDNW is 2,890 and 4,138 respectively. But since there are 40 entries in SECD which include word-forming affixes and combining forms as word items, they are outside discussions. As can be seen from the data in Table 1, the proportion of various word-formation methods and loanwords are the productive way to create new words, while the new meaning and new usage are less productive. Among them, various word-formation methods predominate over the other three to generate English new words. Then loanwords come next, which mainly borrow from French and Latin, while some steal from Japanese, Chinese and Arabic. The increase of loanwords, on the one hand, shows the inclusiveness of English and its constant acceptance of new things. On the other hand, it reflects the frequent communication between countries and world integration. The proportion of old words with new meanings decreases, which indicates that people are more inclined to express new concepts in new ways. This paper mainly focuses on the analysis of various word- formation methods in SECD and ECDNW.

Table 2 Words though Word-formation in SECD and ECDNW

Word- formation	SECD	Proportion in SECD	ECDNW	Proportion in ECDNW
compounding	1427	56.7%	1887	49.7%
Derivation	367	14.6%	538	14.2%
Abbreviation	376	15.0%	525	13.8%
Blending	142	5.6%	393	10.3%
Conversion	61	2.5%	221	5.8%
Others	142	5.6%	231	6.2%

3.1 Compounding

A compound word is a word formed by two or more words put together. As can be seen from Table 2, 1427 and 1887 compound words are included in SECD and ECDNW respectively. Compound words come in many forms, but the most common type is a combination of two simple words. Jackson et al. (2000: 81) argue that compounds with two roots are the simplest and most common type, and that their initial words are more concentrated. This point is also demonstrated in SECD and ECDNW. The head words of compound words in SECD mainly focus on action, air, alpha, alternative, baby, big, body, boom, and so on. The head words in ECDNW mainly focus on carbon, cloud, citizen, dark, data, deep, digital, flash, food and so on. The differences of the head words also have certain characteristics of their special background. For example, the words baby boomer, baby boomerang, baby break, baby bust and baby buster in SECD are the legacy of the ups and downs of the fertility rate in human history. The head words of ECDNW, such as carbon and cloud reflect people's concern for the environment in the 21st century and the concept of computer-related cloud storage in the era of big data respectively.

Table 3 Compounds of Different Word Class in SECD and ECDNW

	Noun	Adjective	Verb	Adverb	Interjection	Total
SECD	1300	115	68	3	1	1427
ECDNW	1661	168	52	2	1	1887

From the perspective of word class classification, it can be seen from Table 3 that there are 1300 nouns, 115 adjectives, 68 verbs, 3 adverbs and 1 interjection in SECD's compounds. There are 1661 nouns, 168 adjectives, 52 verbs, 2 adverbs and 1 interjection in ECDNW's compounds. It can be seen that the proportion of each part of word class in the two dictionaries is similar, in which the dominant part is compound noun, followed by compound adjective, compound verb, compound adverb and compound interjection. In terms of morphology and spelling, compound words can be classified as closed-compound, open-compound and hyphenated compound. It is worth noting that there are 227 compound nouns in SECD, accounting for 15.9% of its compound words. There are 158 compound nouns with hyphens in ECDNW, accounting for 8.3% of its compound words. It can be seen that the use of hyphens in compound words is declining, which reflects the conciseness in word formation. From the perspective of formation, most compound words in SECD and ECDNW contain two words, few compound words formed by three words, and even few compound words formed by four words and over. In terms of compound words formed by the three words, there are some commonalities between the two dictionaries. First, they are usually combined with word and, such as Doom and Gloom, Cap and Trade and Click and Collect. Secondly, they are combined with two hyphens, such as just-in-time, labor- delivery-recovery, and back-to-back.

There is a kind of word formation component in English which is no different from affixes in form. It is characterized by hyphens and can form new words with other words or components. However, the word formation is generally regarded as compound words rather than derivative words like affixes. Such elements are called combining form. Combining form, just like the head word in compound words, has similar background significance. For instance, info-, as a combining form, in SECD is truncated from information. In the nineties of the 20th century, the computer has a great leap in its involution, followed by the overwhelming network information. The compound words created though combining with info- like infobahn, infometric, infonaut, informate, information. In ECDNW, the compound words with info- are infobese, infobesity, infographic, infomania, infomaniac, infocurrency and infomore. Ten years passed, info - as a combining form still has its vitality. Combining forms like cyber- and -ware share same vitality with info-. However, some combining forms such as mega- and -must in SECD have gone through extinction. At the same time, others, such as auto- in ECDNW, have regained their vitality.

Over the past 20 years, the number of new English words composed though compounding is still far more than that the rest of word formation methods. This is partially because compound words are composed of putting two or more words together. Compared with other word-formation methods, compounding is simpler and more suitable for the endless new things and new concepts in today’s world. And the new words constructed by it are obtained by combining the old words, so they are easier to remember.

3.2 Derivation

There is a predictable way of word formation in English Vocabulary: one can create new words by adding affixes at the beginning or end of the source word. This way of word formation is called derivation, and the new word from that is called derivative. Derivation refers to the process of adding affixes to the original words to form a new language. It is also called affixation, which can be divided into prefixation and suffixation.

Table 4 Derivative through Prefixation and Suffixation in SECD and ECDNW

	Prefixation	Proportion	Suffixation	Proportion
SECD	53	14.5%	314	85.5%
ECDNW	144	17%	394	73%

According to Table 4, there are 53 new words derived from prefixes and 314 derived from suffixes in SECD, accounting for 14.5% and 85.5% of its total derivatives respectively. There are 144 new words derived from prefixes and 394 from suffixes in ECDNW, accounting for 17% and 73% of its total derivatives respectively. Two sets of data show that there are far more words derived from suffixes than from prefixes in derivatives. Among them, the common prefixes in SECD are anti-, bi-, de-, e-, re-, super-, trans-, up-; common prefixes in ECDNW are a-, anti-, cis-, co-, de-, gero-, non-, over-, post-, pre-, self-, super-, trans-, UN -. It can be seen that such prefixes as anti-, de-, trans- and super- still have vitality and constantly derive new words after 20 years. It’s worth noting that cis-, an older prefix from Latin meaning “on the other side”, acquired a new meaning in the 1990s refreshed to create new words including cisgender, cisgendered, cisnormative, cisphobia, cissexism, cissexist.

Table 5 Derivative through Suffixation of Different Word Class in SECD and ECDNW

	Noun	Adjective	Verb	Adverb
SECD	230	64	26	4
ECDNW	297	75	19	1

Suffixes, as the word-forming components placed after source words, mainly play the role of changing the word class of the source word. As can be seen from Table 5, the proportion of word class derived from suffixes in the two dictionaries is similar, in which nouns predominate in derivatives, followed by adjectives, verbs and adverbs. Among them, the common noun suffixes in SECD and ECDNW include -ism, -ist and -zation; adjective suffixes like -ed, -ing, -ic, -cal, -er, -ize; verb suffixes like -ize, -ify; adverb suffixes like -ly and -ing. The above suffixes exist facultative and can be used as both nouns and adjectives, or both adjectives and adverbs.

In general, prefixes and suffixes are stable to some extent, even if some new prefixes or suffixes acquire new meanings in a certain historical context with strong derivative ability. And derivation still provides main impetus in the expansion of English vocabulary.

3.3 Abbreviation

Table 6 Acronym in SECD and ECDNW

	Acronym	Proportion
SECD	232	61%
ECDNW	363	69%

There are hundreds of languages in the world. Although they are quite different in many ways, there are some similarities between them. One of them is to simplify words. In English, abbreviation is mainly presented in three forms: abbreviation from, initialism and acronym, and clipping. The common abbreviation form words in SECD and ECDNW include neg (negative), rep (representative), cellphone (cell phone) and so on. In addition, there are 232 acronyms and 363 acronyms in the supplement and the new era, accounting for 61% and 69% of the acronyms respectively, more than half of their abbreviation. The acronyms in the two dictionaries mainly consists of three letters, such as AFK (away from keyboard), ALM (application lifecycle management), APT (advanced persistent thread), then followed by two letters and four letters, such as AR (augmented reality), BD (bond and discipline), ASMR (autonomous sensory meridian response), and BAME (black, Asian or minority ethnic). The form of acronym is usually in capitals. Most clipping words in SECD and ECDNW are nouns and few adjectives.

3.4 Blending

Blending is a word-formation process where a novel word is created from two or more source forms. Originally not productive, it is gaining momentum in modern English. Cannon (1986: 725) took the view that blending is “one of the most intricate of all the word-forming categories”. Algeo (1991: 10) also considered that “this simple process [blending] has a number of variations, some quite complex”. Contemporary blends are on the increase and incorporate new features.

Table 7 Blends of Different Word Class in SECD and ECDNW

	Noun	Adjective	Verb	Adverb
SECD	134	7	4	0
ECDNW	362	17	13	1

As can be seen from Table 7, the blends in SECD and ECDNW are mainly nouns, followed by adjectives and verbs, with the least number of adverbs. In other words, noun blends are major part of blends. The meaning formation of blends does not simply combine the meaning of the source words, and the semantic relations between the source words play a significant part (Gries 2012; Beliaeva 2014). In the noun blends, the first component usually plays a modifying role, while the second component is the decisive factor of word meaning. For instance, in the word "advertgame", the former part of "advertisement" only indicates the type of the latter game, and game determines the meaning of the blended word. Other examples include advertisement, babymoon and webcam. What's more, a sense of banter is harbored in blends. For example, out of ridicule, someone creates such blends as moobs (man + boobs), floordorbe (floor + wardrobe), Bobo (bourgeois + Bohemian). It is also partly for this reason that most blends are used in informal situations.

3.5 Conversion

From the angle of English evolution, conversion has a strong incentive to enrich English vocabulary. Jackson et al. (2000:86) believed that conversion "is a rich source of new words, because it has no restrictions on form".

Table 8 Words through Conversion of Different Word Class in SECD and ECDNW

	Noun	Adjective	Verb	Adverb	Pronoun	Interjection
SECD	10	8	43	2	1	0
ECDNW	79	39	96	6	0	1

It can be seen from Table 8 that most words in SECD and ECDNW are verbs shifted from nouns. The number of noun in ECDNW is more than that in SECD, in which verb is mainly shifted from noun. In the process of conversion, few words are converted into adverbs, pronouns and interjections. It can be seen from Table 2 that compared with other word-formation methods, the weight of conversion in SECD and ECDNW has been boiled down.

4. Conclusion

Through a comparative study of neologisms in SECD and ECDNW, it is found that word formation is the main source of new words, among which compoundingpredominates over derivation, abbreviation, blending and conversion. As for compounding words, the majority is mainly nouns which are often placed two simply words together. Compounds reduce to use hyphen to combine words and its head words have a recorded history of a certain period of time. As for derivation, new words derived from suffixes are far more than those derived from prefixes. The types of prefixes and suffixes are in a stable situation, though some new affixes will regain new meanings in specific contexts, thus having the ability to create new words. As for abbreviation, most of them are acronyms and clipped words are mainly nouns with a small number of adjectives. As for blending, blends are frequently used in a informal situation due to its banter sense, and its core meaning is often decided by the second word if a blend is composed by two words. As for conversion, words are mainly verbs shifted from nouns but the tendency of verbs shifting to nouns is in a rise.

The purpose of this paper is to find out the similarities and differences between SECD and ECDNW by comparing English neologisms, and to sum up some shared rules. It should be pointed out that the number and breadth of the new English words studied in this paper are limited, and the data analysis is not detailed enough, which needs to be improved in the future.

5. References

Adams, Valerie. 2001. *Complex words in English.* Harlow: Longman.

Algeo, J. 1991. *Fifty years among the new words: A dictionary of neologisms, 1941-1991*. Cambridge: CUP.

Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.

Beliaeva, N. 2014. A Study of English Blends: from Structure to Meaning and back Again. *Word structure* 1: 29-54.

Cannon, G. 1979. Affixation and Compounding in New Words in Primarily American English*. Meta*, 24 (3), 326–335.

Cannon, G. 1986. Blends in English Word Formation. *Linguistics* 24: 725-753.

Cao Jianxin. 1994. A Review of Newmark's Translation of Neologisms. *Chinese Translators Journal*, (03): 8-11.

Cheng Ling, Duan ran, Jiang Ying, et al. 2015. Analysis of word formation and translation of English neologisms. *Overseas English*, (1): 59-64

Gao Fengjiang. 2001. The Word-formation and Characteristics of English Neologisms. *China science & technology translators Journal*, (4): 54-45.

Gao Yongwei. 2018. Preface to *The english-Chinese dictionary of new words*. Beijing: The Commercial Press.

Gao Yongwei. 2018. *The English-Chinese dictionary of new words*. Beijing: The Commercial Press.

George. J. Xydopoulos, Irene Lazana. 2014. *A View into retronymy as a source of neology.* Lyon : Publications du CRTT.

Gries, S. T. 2012. Quantitative Corpus Data on Blend Formation: Psycho- and Cognitive Linguistic Perspectives// V. Renner, F. M. Pierre & J. L. Arnaud. *Cross- disciplinary perspectives on lexical blending*. Berlin: De Gruyter Mouton: 145- 167.

Guo Nianzhong. (1990). On the Word-formation of New English Words. *Foreign language education*, (03): 94.

Hao Yue. (2018). A Comparative Analysis of New Word-formation between English and Chinese. *Overseas English*, (24): 89-90.

Jackson, Howard & Etienne Zé Amuela. (2000). *Words, meaning and vocabulary*. London: Cassell & Co.

Ji Shaofeng, Liu Weidong. (2006). On the Method and Skill of Guessing the meaning of English new word's meaning. *Education and Vocation*, (30): 92-94.

Li Lu. 2018. Cognitive Analysis of Word Spy 2016 New Words in Word-formation from the Concept Integration Theory. Harbin: Harbin Engineering University.

Lu Gusun, (1999). *A Supplement to the English-Chinese Dictionary (Unabridged)*. Shanghai: Shanghai Yiwen Press.

Marchand, Hans. (1969). *The categories and types of present-day English word- formation*. 2nd edition, 1st edition. Münich: C.H. Beck.

Mattiello, Elisa. (2017). *Analogy in word-formation : A study of English neologisms and occasionalisms.* Berlin/Boston, Germany : De Gruyter Mouton, *26*.

O’Dell, Felicity. (2016). Creating new words: affixation in neologism. ELT Journal: *English Language Teaching Journal*. 70 (1), 94-99.

Song Ziran, Yang Xiaoping. (2006). *Chinese Neologism Annual Edition (2003-2005)*, Sichuan: Bashu Publishing House.

Wang Rongpei. (2000). The origin and prospect of English neologisms. *Foreign Languages and Their Teaching*, (9): 7-11.

Wang Shuang. (2014). The Trend of Contemporary English Word-formation: An Analysis of New English Words Based on Online Oxford English Dictionary. *Journal of Chifeng College (Chinese Philosophy and Social Sciences Edition)*, 35 (06): 196- 198.

Wang Xiaohan. (2014). *The word-formation and translation skills of new English words in the internet*. Jilin: Jilin University.

Yang Yonglin. (1997). The Expression of Culture in English Neologisms-An Analysis of -gate and its Compounds. *Journal of Northwest Normal University (Social Science)*, (1).

Yu Jing. (2007). *A study of metaphorical word-formation of new English words*. Hebei: Hebei University.

Zhang Yihua. (2003). The Emergence and Structural Motivation of Neologisms in the Information Age. *Lexicographical Studies*, (05).

ENCODING AND PUBLISHING IDIOM DICTIONARIES USING XML TECHNOLOGIES

Ana Rita Vieira[1], Idalete Dias[1], Alberto Simões[2]
[1]Center of Humanistic Studies/ILCH, Universidade do Minho, Braga, Portugal
pg38980@alunos.uminho.pt; idalete@ilch.uminho.pt
[2]2Ai, School of Technology, IPCA, Barcelos, Portugal.
asimoes@ipca.pt

Abstract

This project aims to provide an interesting and efficient way to publish a bilingual lexicographic online resource of idiomatic expressions. In this prototype, we are encoding and making available Schemann's *Synonym Dictionary of German Idioms* and his bilingual *German-Portuguese Idiomatic Dictionary*. These two dictionaries follow two completely different structures, as the *Synonym Dictionary* uses an onomasiological framework, in contrast to the usual semasiological approach applied in the *Bilingual Dictionary*.

These dictionaries were encoded with the Text Encoding Initiative (TEI) schema and its search is supported by the eXist-DB database, following other works on Online Dictionary publishing. Encoding a dictionary of idioms is not trivial, and it gets more complicated when including different information sources. Thus, and while TEI has a comprehensive set of tags to encode dictionaries, it needs adaptations to be able to encode our data properly.

Another challenge for this project is to understand what is the best way to allow the user to search the dictionaries, finding the desired information, focusing on how to allow the proper use of an online onomasiological dictionary. The current prototype allows the users to search for idiomatic expressions by words, by concepts or to browse an ontological structure. This is supported by the cross-reference and linkage of the dictionaries, bringing together the onomasiological and semasiological approaches. Focused on the user's needs and according to the most recent online dictionary studies, the search tool is prepared to help the users in lexical reception and production, as well as in translation tasks.

Keywords: Idiomatic Expressions, Online Dictionary, Text Encoding Initiative, XML, XQuery

1 Introduction

Online bilingual lexicographical resources specialized in idioms are still quite scarce to this day. If managing multi-word expressions is a great challenge, dealing with idioms is an even greater challenge. Both lack the principle of compositionality, as the meaning of the whole is not obtainable from the meaning of its constituent parts. But while it is rather easy to find common multi-word expressions in dictionaries, lexicons, and computational corpora, there is a significant lack of resources containing idioms.

This project aims to fill in this gap by creating electronically encoded idiom dictionaries based on dictionaries available in print. In this sense, the *German-Portuguese Idiomatic Dictionary*, a semasiological dictionary, and the *Synonym Dictionary of German Idioms*, an onomasiological dictionary, both compiled by Hans Schemann, provided us with the required information. As studies on the modelling of online lexicographic data (Klosa, 2013; Müller-Spitzer, 2014a) have shown, the process of computerizing and publishing dictionaries on the Internet is not a trivial task, since the relationship between the dictionary data, specific user needs, and possible access routes and search paths to satisfy user demands must be considered (Tarp, 2013). One of our objectives is to use all the information available in the above-mentioned print dictionaries, including structural features and content, to design an integrated model capable of enabling users to carry out focused searches. To achieve this, a detailed XML schema was developed, on the one hand, to represent the complex system of lexicographic information contained in the dictionaries and, on the other, to support the computational tool that will facilitate the dictionary use.

In this paper, we describe the modelling and encoding principles applied to the dictionaries with a focus on: (i) integrating different lexicographic structures and types of information in one electronic dictionary; (ii) providing the users with useful results that match queries in specific communicative and cognitive situations. Therefore, as a first step, we were faced with the challenge of designing a proper granular annotation schema using a dictionary encoding standard to represent both dictionaries' macro and microstructures at the same time ensuring the interconnectedness between the dictionaries' entries and the modelling of all possible querying outputs.

Following the encoding process, the next challenge was to create a search interface that can, on the one hand, mirror and preserve the dictionaries' semasiological and onomasiological approaches; and, on the other hand, provide easier and quicker methods to respond to the needs of the target users. Every decision taken throughout this project was taken with the user in mind, with well-defined target users, user needs, and user lexicographical situations, as advocated by the Function Theory of Lexicography (Tarp, 2012, 2013, 2014) combined with the most recent studies on online dictionary use.

The current prototype is best suited for Portuguese advanced German L2-Learners and German advanced Portuguese L2-Learners, as well as translators, guiding them throughout the interface according to their communicative situations of text reception, text production, and translation.

The remainder of this document is structured as follows: Section 2 discusses related work, focusing on different projects that are publishing dictionaries online, including idiom dictionaries. Section 3 describes the characteristics of the two print dictionaries that were annotated for this project. The annotation schema is explained in Section 4, and Section 5 focuses on the prototype developed to support the user experience. Finally, we conclude in Section 6 with some insights on the current project status and the expected future developments.

2 Related Work

In the last decade, we have witnessed a clear decline in the creation and selling of printed dictionaries. With the advent of digital devices and continuous access to the Internet, users tend to use online tools to gather information about words and expressions. Recent studies on user behaviour have shown that searches on a language problem are performed directly in a web search engine and not in a dictionary website (Sascha et al., 2018). The result is that publishing houses are not interested in further developing their dictionaries, and have been focusing on making them available on the Internet. We can find examples of online dictionaries for most languages, from major publishers: for Portuguese (*Porto Editora*[1]), English (*Oxford*[2] and *Cambridge*[3] Dictionaries), French (*Larousse*[4]), or German (*Duden*[5]), just to mention a few. The same is true for Academy dictionaries, like *Real Academia Española*[6], *Académie Française*[7] or *Academia das Ciências de Lisboa* (Salgado et al, 2019).

[1] Integrated in the Infopedia Web Portal, https://www.infopedia.pt/.
[2] Available at https://www.oed.com/, through subscription
[3] Available at https://dictionary.cambridge.org/.
[4] Different dictionaries available at https://www.larousse.fr/.
[5] Available at https://www.duden.de/.
[6] Available at https://dle.rae.es/.
[7] Available at https://www.dictionnaire-academie.fr/.

Most of this work still follows a paper-based approach (Tarp, 2012): each entry of the dictionary is stored in a database, allowing the user to search for the headword and, in some specific situations, by words present in the definitions. The relationship between entries is done in the same way as was presented on paper, with specific "*see also*" sections. More recently, as there is a large diversity of dictionaries online for every language, some attention is being given to the interaction with the user, not just by giving them cleaner interfaces, but also presenting more information than just the word entry.

When it comes to onomasiological dictionaries, most of the available projects on the Internet are thesauri. Some examples are the *Old English Thesaurus*[8] or *VisuWord*[9] for the English Language, the *Caldas Aulete* dictionary for Portuguese[10], the *OpenThesaurus* for German[11] or the multilingual *ConceptNet*[12]. These network-like dictionaries can be compared to the diverse *WordNet* (Miller, 1995) projects, available for most languages.

There are three other lexical projects that we would like to emphasize for supporting both semasiological and onomasiological queries: The *ANW Dictionary* (*Algemeen Nederlands Woordenboek*) [General Dutch Dictionary], *CombiDigiLex* and *Tesouro do léxico patrimonial galego e português* [Galician and Portuguese word bank]:

- The *ANW Dictionary*[13] is "not a clone of an existing printed dictionary [and] it truly represents a new generation of electronic dictionaries in the sector of academic and scientific lexicography" (Moerdijk, 2008). It is a corpus-based dictionary of written Dutch and it pays special attention to '*semagrams'*: "conceptual structure elements which characterise the properties and relations of the semantic class of a word meaning", playing an important role in onomasiological queries (idem, 2008).
- *CombiDigiLex*[14] is an ongoing multilingual project, focused on the analysis of the possible lexical combination of 'communication, movement, emotion, perception and transfer verbs' in German, Spanish and Portuguese, framed within a conceptual macrostructure and built to assist in L2-text production;
- *Tesouro do léxico patrimonial galego e português*[15] is a lexical open data portal for the Galician, European Portuguese and Brazilian Portuguese lexicon. Querying this database provides information to be of use for dialect comparative studies, presenting the search results by a choice of lexical forms, by location and by semantic field. The search interface is built around a semantic classification of words and idioms, composed of twelve major categories with further divisions. To search by concept, the user has to go to the advanced search tool and select a concept (eg. 7.1 - Humans [physical, psychological and behavioural aspects]), showing a list of results that can be filtered by location and grammatical category, for instance selecting 'adverbial expression' (eg. *em leitão* [naked], an European Portuguese adjectival and adverbial expression, which is also an idiom).

Regarding Idiom dictionaries, there are relatively few projects on the web. Most of the available projects are not based on academic dictionaries normally associated with a publishing house. Here we will refer to a few:

[8] Available at https://oldenglishthesaurus.arts.gla.ac.uk/category/.

[9] Available at https://visuwords.com/.

[10] Available at https://www.aulete.com.br/analogico/.

[11] Available at https://www.openthesaurus.de/.

[12] Available at https://conceptnet.io/.

[13] Available a https://anw.ivdnt.org/.

[14] See more at https://combidigilex.wixsite.com/website.

[15] Available at http://ilg.usc.es/tesouro/en.

- *The Idioms*[16] is focused on the English language. The idiomatic expressions in this resource are organized by topics that lead the user to a list of idioms that are related to the chosen topic. As an example, clicking on the topic 'problem' will present the user with a list of entries, each containing the idiomatic expression (*time puts everything in its place*; *pour oil on troubled waters*; *elephant in the room*; *a hard nut to crack*, etc.), the definition and an example sentence. The user can decide if more information should be presented on a specific idiom by clicking on the button 'Read on': the user will have access to paraphrases, more example sentences, information on the origin of the expression and synonyms. A second access route is provided by the 'Complete List' option that lists all the idioms in the database, although it is not quite clear which ordering strategy was followed. The interface allows for a third access route via a search field. All expressions and example sentences that contain the word entered in the field will appear.
- *Dictionnaire d'Expressions Idiomatiques*[17] is a bilingual Portuguese - French resource of idiomatic expressions. One may search the dictionary via a given list of the expressions or concepts in both languages, resource which serves as a great reference to our project of an online dictionary featuring both semasiological and onomasiological approaches and also synonym relations.
- *Expressio*[18] includes monolingual, bilingual and multilingual dictionaries of idiomatic expressions. When using the multilingual version, results are shown in a table, presenting the equivalent expression in the target language along with its literal translation. For each entry, it includes examples and suggests other idioms with similar meaning. Users can comment on the entries, and suggest changes in the dictionary. Taking the bilingual Portuguese - French as example, for each expression, the user has access to the variant of Portuguese (European or Brazilian) to which the idiom belongs, the French equivalent expression, and a literal translation of the Portuguese source idiom in French, a feature that could prove useful for French learners of Portuguese as a foreign language. For example, the search results for the Portuguese idiom "*não se dar por vencido*" [*to keep one's chin up*] will include "*contre mauvaise fortune bon coeur*" (French equivalent) and "*ne pas se donner comme vaincu*" (literal translation) [do not give oneself as defeated].

3 Dictionaries/Resources

Before initiating the electronic encoding of the dictionaries, we needed to understand how the idioms are presented and organized in both dictionaries and what are the different approaches to using them, particularly in a situation in which the user has to search in both dictionaries to find what he/she wants. The user is faced with a challenging task due to the differences in the dictionaries' structures, as the *Synonym Dictionary* uses an onomasiological approach, in contrast to the usual semasiological approach applied in the *Bilingual Dictionary*. In the former, the idioms are grouped into synonym sets (synsets) which, in turn, are organized into concepts, according to Schemann's knowledge categorization, creating a hierarchical structure that allows the user to explore synonymous expressions that convey the same concept. This lexicographic perspective is particularly useful for writing tasks, that is encoding purposes, to assist users who are looking for expressions that designate a specific concept in their native or a foreign language (Sierra, 2000).

The *Bilingual Dictionary*, on the other hand, follows the more usual alphabetical macrostructure. However, this ordering principle may also present some difficulties to the user because idioms are multi-

[16] Available at https://www.theidioms.com/.

[17] Available at https://www.cnrtl.fr/dictionnaires/expressions_idiomatiques/.

[18] Available at https://www.expressio.fr, https://www.expressio.fr/expressions-idiomatiques-en-portugais.

word expressions. The user will be faced with the challenge of figuring out which keyword was used by the lexicographer to catalogue the idiom. In addition, whenever the same expression is catalogued under different entries, there will be cross-references instructing the user to search for a related entry where the full entry content data can be consulted. This feature may contribute to frustrating user experiences. Note that we are dealing with printed editions, and therefore the lexicographer is bound to space constraints demanding ingenious ways to plan and compile the different textual structures: from the data distribution structure, the access structure, the macrostructure, microstructure, mesostructure, not to mention the addressing structure and search zone structure (Wiegand, 1990; Müller-Spitzer, 2014a). Despite the detailed description of the ordering strategies provided by Schemann in the outer text of the *Bilingual Dictionary*, there is general agreement that, due to several reasons, including the lack of a dictionary culture which can be achieved by dictionary pedagogy, users seldom consult(ed) the outer texts for guidance on how to use a print dictionary (Gouws, 2010). For electronic and online dictionaries we are not dependent on the number of pages allowing the use of technology to automate these indirections and offer a better user experience.

3.1 *Synonym Dictionary of German Idioms*

The macrostructure of the *Synonym Dictionary of German Idioms* was obtained by applying a bottom-up categorization approach that consisted, firstly, in grouping the 18959 German idioms into synsets, followed by the categorization of the synsets with a common denominator into subconcepts that are linguistically realized by what Schemann designates as archilexemes. It is important to emphasize that it is not always possible to find one lexical unit that satisfactorily delimits the semantic content of a group of synsets that belong together. For this reason, Schemann (2012) made use of more than one lexical unit and/or specific idioms within the synsets to act as delimiters. Once at the level of the archilexemes, Schemann grouped these subconcepts into higher-order generic concepts and these, in turn, into nine macro concepts which represent a given organization of the world. Schemann's conceptual system has proved to be a reference for the conceptualization of onomasiological dictionaries (Dias, 2010).

The first section of the dictionary is the browsing section where the nine macroconcepts are presented as follows:

A: *Zeit, Raum, Bewegung, Sinnesdaten* [Time, Space, Movement, Sensory data]

B: *Leben - Tod* [Life - Death]

C: *Physiognomie des Menschen* [Human physiognomy]

D: *Stellung zur Welt* [Attitude to the world]

E: *Haltung zu den Mitmenschen* [Attitude towards fellow human beings]

F: *Einfluß, Macht, Verfügung, Besitz* [Influence, Power, Disposition, Possession]

G: *Kritische Lage, Gefahr, Auseinandersetzung* [Critical situation, Danger, Conflict]

H: *Präferenzen* [Preferences]

I: *Quantitäten, Qualitäten, Relationen* [Quantities, Qualities, Relations]

These concepts are further subdivided into more specific ones, composing a structure made out of three levels, that are identified by a sequence of characters. For instance, the path "*B - Ba - Ba1*" refers to the top-level concept "*B: Leben - Tod*" [Life - Death], the second-level concept "*Ba: Geburt - Tod*" [Birth - Death], and the third-level concept "*Ba1: Geburt*" [Birth]. After selecting the desired concept, the user has to search for it in the second and main part of the dictionary, the *Systematischer Teil* [Conceptual Part], which is where we find the idiom synsets: a collection of idioms grouped under a specific concept, such as

"*Ba1: Geburt*" [Birth] (36 idioms grouped under 21 synsets). The Conceptual Part of the dictionary assists users with text production in the native language and the foreign language. Two possible communication-oriented functions of the dictionary can be described as follows: a native speaker of German who would like to explore possible synonymous idioms that express the same concept by analyzing and comparing the underlying images of the expressions that belong to the same synset and then decide which better fits a given context; an advanced learner of German as a Foreign Language may use the dictionary in a similar situation. In this case, the user might need to resort to complementary lexicographic resources, such as a bilingual dictionary.

The *Alphabetischer Teil* [alphabetical part] contains all the expressions present in the conceptual section of the dictionary, ordered alphabetically by the keyword that was chosen by the lexicographer. Each expression is linked directly to the conceptual part via a synset code, that remits the user to the specific synset in which it is found. This part is more directed to those who have a (key)word in mind and would like to consult the exact form of the idiom. For example, "*(ein Kind) zur Welt bringen*" [*to bring a child/... into the world*] - Ba 1.4. This section was important for the electronic encoding of the dictionary, as it allowed the alignment between the *Synonym Dictionary* and the *Bilingual Dictionary*, as will be described in Section 4.

The last part of the *Synonym Dictionary* is the *Such-und Stichwortregister* [keyword search index], which lists the archilexemes on the subconcept level in alphabetical order. This index at the end of the dictionary forms part of the onomasiological approach as it provides the user with another way to search for a concept via the archilexemes or keywords that are used to categorize the synsets.

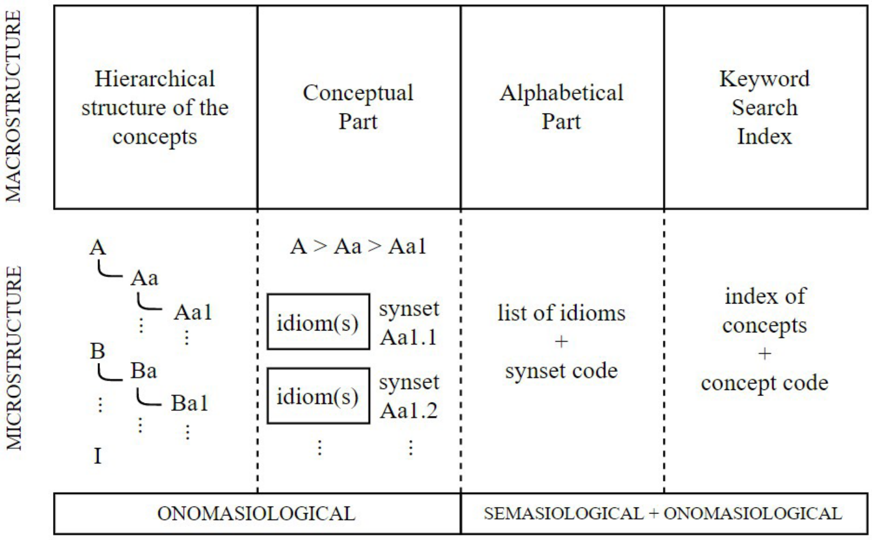


Figure 1 - The *Synonym Dictionary*'s macro and microstructures.

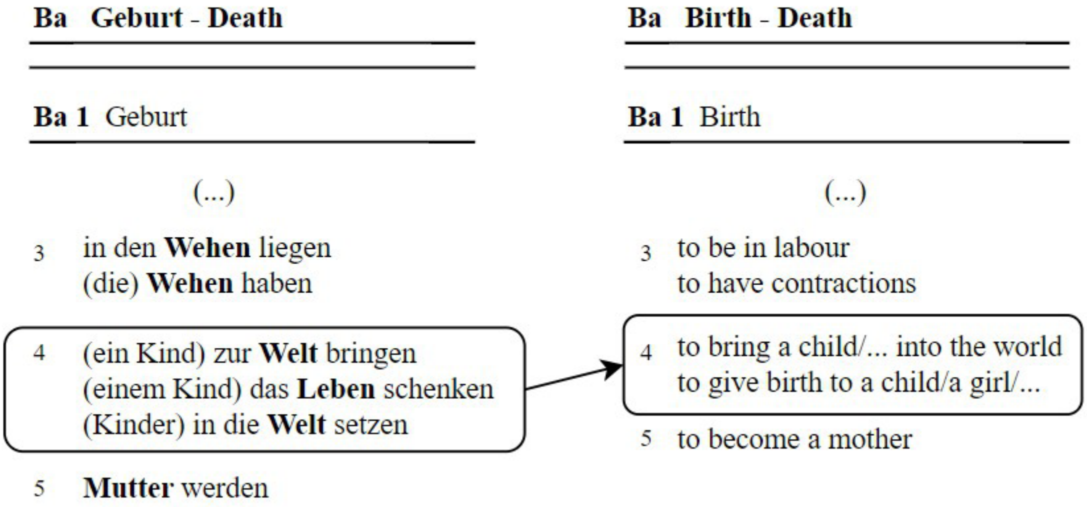
To sum up, the users can search this dictionary starting from any one of the sections presented in Figure 1. The sections can be chosen according to the users' knowledge of the language, the usage situation and corresponding cognitive and communication-oriented functions. The functions are presented in more detail below:

(i) Cognitive functions:

- The user is interested in consulting Schemann's categorization of the world and the methodology used to construct such a system. The user might also be interested in comparing Schemann's conceptual system with other systems that have been developed, such as the *Diccionario Ideológico de la Lengua Española* by Casares, or even take Schemann's system as a reference for a new proposal.

(ii) Communicative function mostly related to text production by a native speaker of German or an advanced learner of German as a Foreign Language:

- The user is familiar with a specific idiom, such as "*(ein Kind) zur Welt bringen*" [*to give birth to (a child)*] and would like to find synonymous expressions. He/she would start by searching in the Alphabetical Part, first by looking for the keyword 'Welt' and, then, selecting the expression from the list of other expressions with the same keyword. The user is redirected to the Conceptual Part through a synset code Ba 1.4, corresponding to the fourth synset belonging to the third level concept *Geburt* [Birth]. In this synset, the user will find the idiom he/she started from and two other synonyms: *(einem Kind) das Leben schenken; (Kinder) in die Welt setzen* [*to bring a child/... into the world, to give birth to a child/a girl/...*].
- The user would like to convey a specific concept but is not familiar with idioms that are used to express that same concept. In this situation, the Keyword Search Index will assist the user by redirecting him/her to the respective synset in the Conceptual Part. To be able to use the Keyword Search Index, the user will have to think of a clue word related to the concept. Since the keywords in the Index are the archilexemes defined by the lexicographer, the user's clue words may not always match the indexed keywords. The concept of 'birth' is delimited or indexed by the archilexeme 'Geburt' [Birth] in the dictionary. This means that the user will find the information he/she is looking for if the clue word is *Geburt*.
- The user looking for an idiomatic expression that conveys a specific concept may start by browsing the hierarchical concept system and select one of the nine macroconcepts (first-level concepts), followed by the respective second-level and third-level concepts. Once at the level of the synsets, the user will need to go through the synsets grouped under a specific concept to be able to identify which one of the synsets contains the expressions with the nuances he/she is looking for. In this case, the search proceeds from the hierarchical structure of the concept system to the Conceptual Part of the dictionary with the synsets. Figure 2 shows three out of a total of twenty-one synsets that are grouped under the third-order concept *Geburt*.



Ba Geburt - Death

Ba 1 Geburt

(...)

3 in den **Wehen** liegen
(die) **Wehen** haben

4 (ein Kind) zur **Welt** bringen
(einem Kind) das **Leben** schenken
(Kinder) in die **Welt** setzen

5 **Mutter** werden

Ba Birth - Death

Ba 1 Birth

(...)

3 to be in labour
to have contractions

4 to bring a child/... into the world
to give birth to a child/a girl/...

5 to become a mother

Figure 2 - Synset Ba1.4 in the Conceptual Part and its equivalent synset in English[19].

[19] Note that at the present moment we are only using the *Synonym Dictionary of German Idioms*. Nevertheless, we present the same synset in English extracted from the *German-English Idiom Dictionary* from the same author for easier reading.

An analysis of the semantic relation between the synsets in Figure 2 reveals that the synsets are closely related to each other. This exercise of exploring and comparing the subtle semantic differences between the expressions in the contiguous synsets is demanding and can only be achieved by users with an advanced proficiency level. The more familiar one becomes with this dictionary, the more useful it will become.

3.2 *The German-Portuguese Idiomatic Dictionary*

The *German-Portuguese Idiomatic Dictionary*, including about 32.000 German idioms, is one of a series of five bilingual idiom dictionaries compiled by Schemann that stems from the main dictionary *Deutsche Idiomatik* (1993). Although its construction is semasiological (compare Figure 1 with Figure 3), it presents a certain level of complexity. Whereas the *Synonym Dictionary* provides the user with at least three access routes to the main onomasiological part of the dictionary with the synsets, the *German-Portuguese Dictionary* provides the user with only one access route to the entries via the keyword chosen by Schemann to serve as ordering principle. These keywords appear as headwords under which all expressions with the same keyword are grouped. For example, the headword *Welt* [world] subsumes a large number of expressions with the same keyword: *nicht die Welt sein* [*it isn't all that much/long/...*], *das Theater/Bücher/... ist/sind meine/deine/...Welt* [*the theatre/books/... is/are my/her/... world*], *um alles in der Welt* [*at all costs; come what may; for God's sake; for goodness' sake*], including *ein Kind/ein Mädchen/... zur Welt bringen* [*to give birth to a child/a girl/...; to bring a child/... into the world*].

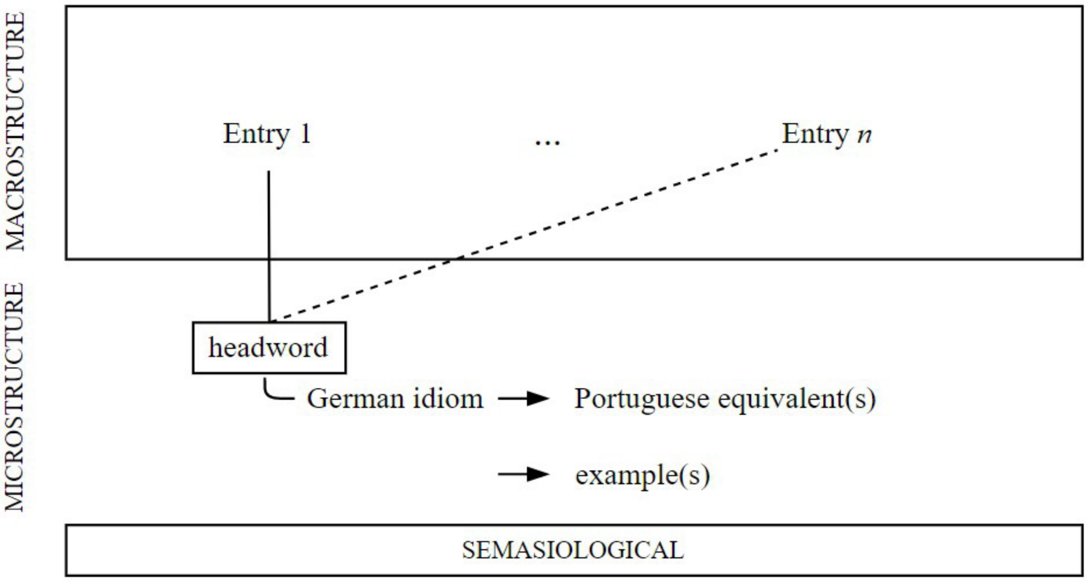
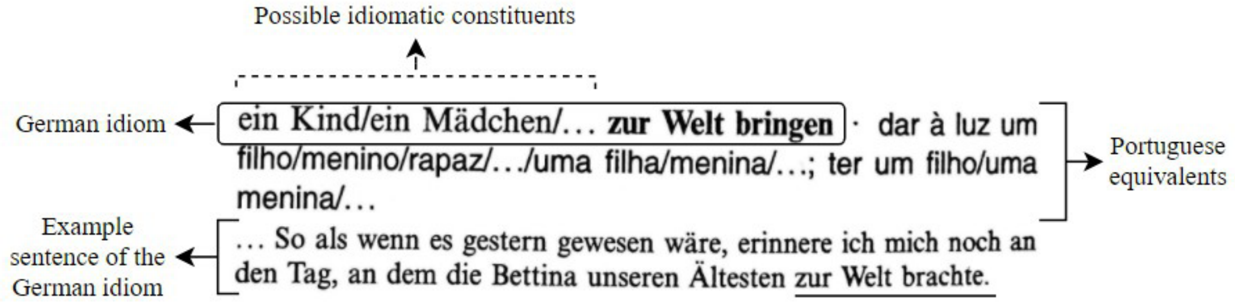


Figure 3 - The *German-Portuguese Idiomatic Dictionary*'s macro and microstructure.

By default, an entry in the *Bilingual Dictionary* has a specific German idiom followed by its translation equivalent(s) in Portuguese and at least one example sentence of the German idiom in context (see example below in Figure 4). To be able to comply with space constraints and to avoid idiom repetition, cross-references appear within the dictionary's microstructure. Therefore, the user is frequently redirected to other entries, which is a common scenario within bilingual dictionaries.



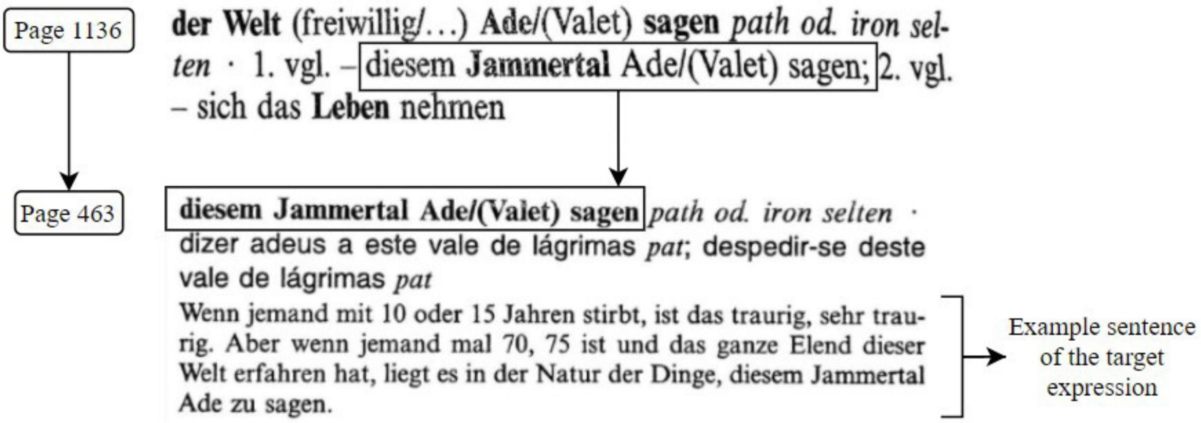
ein Kind/ein Mädchen/… **zur Welt bringen** · dar à luz um filho/menino/rapaz/…/uma filha/menina/…; ter um filho/uma menina/…

… So als wenn es gestern gewesen wäre, erinnere ich mich noch an den Tag, an dem die Bettina unseren Ältesten zur Welt brachte.

Possible idiomatic constituents
German idiom
Portuguese equivalents
Example sentence of the German idiom

Figure 4 - The *German-Portuguese Idiomatic Dictionary's* entry for *ein Kind/ein Mädchen/... zur Welt bringen*.

However, in this dictionary cross-references are used to establish a relation between idioms but also between example sentences. Usage examples are crucial in any dictionary but are of major importance in such a specialized dictionary because they provide a contextualized use of the idioms, which assist in lexical reception and production, as well as in translation. The cross-references in this resource present one main drawback, namely that the act of remitting from one expression (source) to a quasi-synonymous expression (target) leaves the entry of the source expression devoid of an example sentence (Figure 5). The user only has access to the example sentence(s) of the target expression.



der Welt (freiwillig/…) **Ade/**(Valet) **sagen** *path od. iron selten* · 1. vgl. – diesem **Jammertal** Ade/(Valet) sagen; 2. vgl. – sich das **Leben** nehmen

diesem Jammertal Ade/(Valet) sagen *path od. iron selten* · dizer adeus a este vale de lágrimas *pat*; despedir-se deste vale de lágrimas *pat*

Wenn jemand mit 10 oder 15 Jahren stirbt, ist das traurig, sehr traurig. Aber wenn jemand mal 70, 75 ist und das ganze Elend dieser Welt erfahren hat, liegt es in der Natur der Dinge, diesem Jammertal Ade zu sagen.

Figure 5 - The entry for *der Welt (freiwillig/...) Ade/(Valet) sagen* with cross-reference to synonymous expressions

Figure 5 shows the methodology used by Schemann for cross-references. The source entry *der Welt (freiwillig/...) Ade/(Valet) sagen* redirects the user to two quasi-synonymous expressions: 1. *diesem Jammertal Ade/(Valet) sagen [to leave/to say farewell to/to say adieu to/to bid adieu to/...this vale of tears*]; 2. *sich das Leben nehmen* [*to take one's life; to commit suicide*]. There is no example sentence for the source expression, only for the target expressions.

In view of the above, merging the *Synonymy Dictionary* and the *German-Portuguese Dictionary* into one electronic resource taking into account possible ways of interlinking the different contents and modelling common access structures to the lexicographic data will make it possible to cater for a wider range of user types and use situations. In addition, it has been possible to pick up on expressions that belong to a specific synset but were not originally included in the *Synonym Dictionary*.

4 Dictionary Encoding

This section briefly describes the preliminary electronic processing tasks undertaken to merge the two dictionaries, followed by the semi-automatic encoding of the lexicographic data using a subset of the

Text Encoding Initiative (TEI) schema for dictionaries. It also identifies the encoding challenges when annotating these dictionaries.

4.1 Preliminary Electronic Processing

This project emerges after previous work on Schemann's dictionaries[20], which produced the first digital version of the dictionaries in XML files with minimal annotation. The first processing step involved converting the data marked up using TUSTEP (Tuebingen System of Text Processing Tools) into XML with minimal descriptive tags. The second step was dedicated to linking the data between the dictionaries: a Perl script was created to match the German idioms in the *Synonym Dictionary* with the German idioms in the *German-Portuguese Dictionary.* With every match, the synset code in the synonym dictionary was added to the respective bilingual dictionary entry. In order to be able to automatically pick up expressions that were not perfect matches due to differences in paradigmatic lexical units, specific patterns were identified to assist in the matching process. Since the bilingual dictionary consists of almost twice the number of idioms in the synonym dictionary, many German idioms do not have a synset code. One very important result of this process is that the Portuguese equivalents have also automatically been assigned a synset code.

Figure 6 shows the first entry of the *German-Portuguese Idiom Dictionary* in XML with minimal annotation, linked to a synset code from the *Synonym Dictionary*.

```
<eintrag>

   <st>A</st>

   <de><b>das A anschlagen/angeben</b> <i>Musik</i></de>

   <po>dar o lá</po>

   <bs>Mein Gott, ist die Geige verstimmt. Es scheint, du hast nicht das A angeschlagen,
   sondern das H!</bs>

   <kat> Dc10.22 </kat>

</eintrag>
```

Figure 6 - First entry of the *German-Portuguese Idiom Dictionary* with minimal XML annotation.

A minimal entry, `<eintrag>`, included a headword, `<st>` (*Stichwort*), the German idiom annotated with the `<de>` (*deutsch*) element, highlighted using the bold element `<b>.` The Portuguese equivalents were encoded with `<po>` (*portugiesisch*) tag and the example sentence using the `<bs>` (*Beispiel*) element. There are other elements like notes, encoded using `<note>`, and usages (using the italic tag `<i>`) and other non-descriptive markup (e.g. #s for letter-spacing). The element `<kat>` (*Kategorie*) provided the synset code which enabled linking the dictionaries using eXtensible Stylesheet Transformations (XSLT)[21].

[20] The processing tasks were performed by Idalete Dias, Center of Humanistic Studies, University of Minho, and José João Almeida, Department of Informatics, University of Minho.

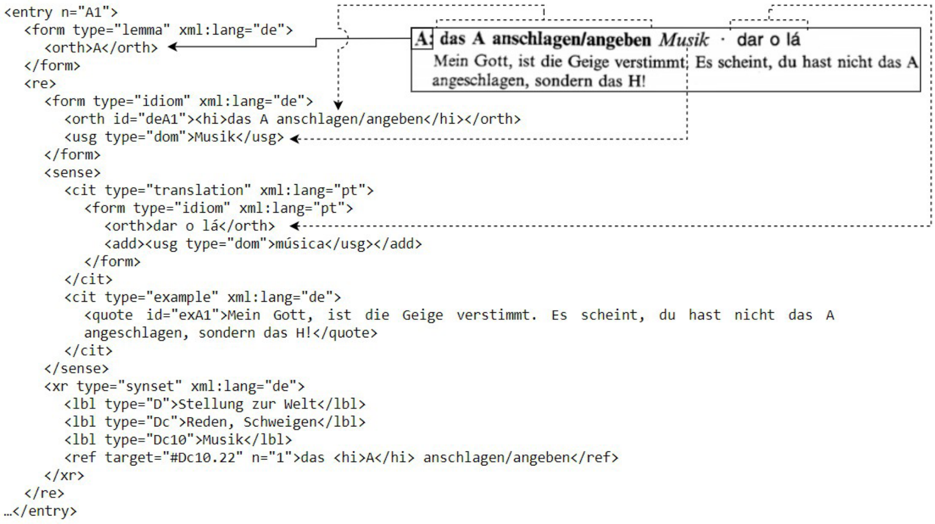
[21] XSLT is a World Wide Web Consortium Recommendation, that defines a language to manipulate and perform structural transformations on XML documents: https://www.w3.org/TR/xslt-30/

4.2 Reencoding into TEI

The Text Encoding Initiative (TEI, 2021) is the result of the work of a consortium of individuals in the creation of an XML schema for Digital Humanities. It supports the encoding of diverse types of resources, ranging from simple prose or poetry up to dictionaries, mathematics, or even linguistic corpora.

While there are diverse formats to encode dictionaries, Chapter 9 of the TEI-P5 schema is one of the most used and was chosen for this project. The conversion of the original encoding into TEI was performed with a set of handwritten rules, compiled in a GNU Make makefile. This allows the use of the rules over the different documents, and to easily edit the encoding process to test new rules.

The selected TEI elements, presented in Figure 7, are suitable for modelling the structure of the dictionary entries and for describing the semantic nature of the components. As long as it is accurate and relevant, the more granular the annotation we apply to the data, the more information the user will be able to retrieve from a given search query.



A: **das A anschlagen/angeben** *Musik* · dar o lá
Mein Gott, ist die Geige verstimmt! Es scheint, du hast nicht das A angeschlagen, sondern das H!

```
<entry n="A1">
  <form type="lemma" xml:lang="de">
    <orth>A</orth>
  </form>
  <re>
    <form type="idiom" xml:lang="de">
      <orth id="deA1"><hi>das A anschlagen/angeben</hi></orth>
      <usg type="dom">Musik</usg>
    </form>
    <sense>
      <cit type="translation" xml:lang="pt">
        <form type="idiom" xml:lang="pt">
          <orth>dar o lá</orth>
          <add><usg type="dom">música</usg></add>
        </form>
      </cit>
      <cit type="example" xml:lang="de">
        <quote id="exA1">Mein Gott, ist die Geige verstimmt. Es scheint, du hast nicht das A
        angeschlagen, sondern das H!</quote>
      </cit>
    </sense>
    <xr type="synset" xml:lang="de">
      <lbl type="D">Stellung zur Welt</lbl>
      <lbl type="Dc">Reden, Schweigen</lbl>
      <lbl type="Dc10">Musik</lbl>
      <ref target="#Dc10.22" n="1">das <hi>A</hi> anschlagen/angeben</ref>
    </xr>
  </re>
…</entry>
```

Figure 7 - The first entry of the dictionary in its print version, compared to its annotation using XML/TEI after linking the dictionaries.

Given the specificity of the document, the TEI schema was extended with some custom annotations to suit the dictionaries' content. For instance, as seen in the example above, the Portuguese idiom "*dar o lá*" [*to hit A*] could also have its usage displayed, which happens to be the same as for its German equivalent. Although the following TEI elements are not allowed as proposed below by TEI's dictionary schema, it is our understanding that these rules best represent the structure and semantics of the dictionary entries:

- <add> must contain the citation element <cit>, in cases where we want to improve the dictionary's content by adding more examples;
- <add> must contain <xr>, when it is clear that a cross-reference is missing;
- <add> must contain <usg>, if an accurate and relevant usage is identified to provide more information to an idiom or an example;
- <cit> must contain <del> and <del> must contain <form>, to delete some clear idiom repetition within the Portuguese equivalents;
- <quote> must contain <usg>, where examples also provide usage information.

These extended specifications allowed us to preserve the dictionaries' original content and distinguish it from new information that was added in this phase of the project. Regarding the last addition, it is of utmost importance to show usage information because an idiom may have several meanings and the nuances between them are given through contextualized examples of their usage situation (e.g. formal, ironic).

4.3 Identified challenges

Two particular challenges emerge from the cross-reference notation in the *Bilingual Dictionary*, one being, as already mentioned, that sometimes an idiom lacks an example sentence because the user is redirected to another entry; and secondly, the user may find more or fewer idioms that share synonymy relationships while going through the cross-references, meaning that, depending on the starting point, the user may never come across some idioms because this linking system is not bidirectional. Furthermore, German idioms that only appear in the *Synonym Dictionary* do not have example sentences because there were idioms that did not match with the ones in the *Bilingual Dictionary* due to slight differences (e.g. punctuation or other suggested co-occurrence lexical units) or German idioms that have no Portuguese equivalents. These issues stem from the print version and we are working to improve this with the aim of providing all idioms with example sentences and Portuguese equivalents.

Further challenges appeared after the TEI encoding of the dictionaries, namely, idioms that point to a synset without belonging to it, when they should; and, sometimes, two or more idioms appearing together separated by a slash, that may not be clear to the user. Therefore, we have identified further annotation improvements within the idioms to provide better assistance in idiom learning.

In regard to the search interface, querying by concept should be improved by implementing an autocomplete function to the search box. If the user chooses to type in the search box, rather than explore the semantic field structure to check for the available concepts, the query may return no results, which is unproductive and frustrating for the user. Suggesting keywords in the search box will not only assist in writing without spelling errors but also instantly show which words are identified as concepts according to the *Synonym Dictionary*.

5 Web Application

In this section, we briefly describe the technologies that are supporting our prototype and share our ideas on how to develop an intuitive and versatile interface for querying an Idiom Dictionary.

5.1 Supporting Technologies

Considering the choice to encode the dictionary using XML technology (namely using the TEI schema), the prototype was developed with technology focused on the World Wide Web Consortium (W3C) standards, namely supporting XPath and XQuery for the querying of the XML dictionary files, and HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript (JS) for building the interface.

With this goal in mind, a set of different document-oriented databases were analyzed, and eXist-DB was chosen. eXist-DB is an open-source document-oriented database that treats XML documents as first-class citizens in its environment. This means that the encoded files can be imported directly into the database, without any kind of extra-processing, and that all the database environment is prepared to deal with such documents.

eXist-DB works not just as a database, but as a full development framework. It allows the development of web applications with XQuery and supports direct XML transformations through eXtensible Style Sheets Transformations (XSLT).

Thus, the XML documents encoded in TEI (one per letter of the alphabet) were imported into the database, and the full interface, discussed in the next section, was developed using XQuery and web technologies (HTML, CSS, JS).

5.2 Prototype Development Guidelines

The creation of an online dictionary requires well-thought considerations related to its "content, presentation, users and usage" (Klosa, 2013), hand in hand with Tarp's lexicographical function theory (2014), which states the following:

- *Dictionaries are utility tools*
- *designed for consultation*
- *and produced with the genuine purpose of meeting punctual information needs,*
- *which specific types of potential user*
- *may have in specific types of extra-lexicographic situation,*
- *by providing access to carefully prepared data*
- *from which the users can retrieve information*
- *which can subsequently be used for different purposes.*

Therefore, the search interface must focus on the targeted users - German and/or Portuguese advanced learners and translators - and their information needs.

Given the dictionaries' content, we have identified the user's needs in communicative situations and cognitive situations. 'Communicative situations' refer to "text reception and production in the mother tongue or a foreign language, translation from and into the mother tongue, and text revision" (Tarp, 2014), while 'cognitive situations' occur when there is a need for knowledge acquisition. The situations covered by the search tool range from Portuguese text production, German-Portuguese translation, Portuguese-German translation, German text reception and text production and Portuguese text reception to users with German or Portuguese as their mother tongue. This makes a total of 12 identified lexicographical situations for which we developed the 'search methods' for idiom comprehension, assistance in writing and translation, as well as the advanced search for more focused results.

In regard to text reception needs (Leroyer, 2018), even though we do not have a definition of the idioms, we can still assist in idiom comprehension (also a cognitive situation) by providing the idiom's synset, whenever it is possible, and/or one or more contextualized examples.

Though the search tool provides results for each of the use situations listed above, it's mostly directed towards tasks related to text production, which can also involve translation and idiom comprehension (Fuertes-Olivera & Bergenholtz, 2018). While full-text search is always provided and the user may search for one or more words that are part of an idiom or identified as a concept, he/she can also explore the hierarchical structure of the concepts, performing an onomasiological search that mirrors the *Synonym Dictionary*'s approach, as intended. This is particularly useful if one is looking for idioms related to a specific concept for writing assistance or, for instance, looking for synonyms while performing text revision.

The search interface also assists in translation from German to Portuguese and its reverse, providing all the information available related to an idiom in both languages.

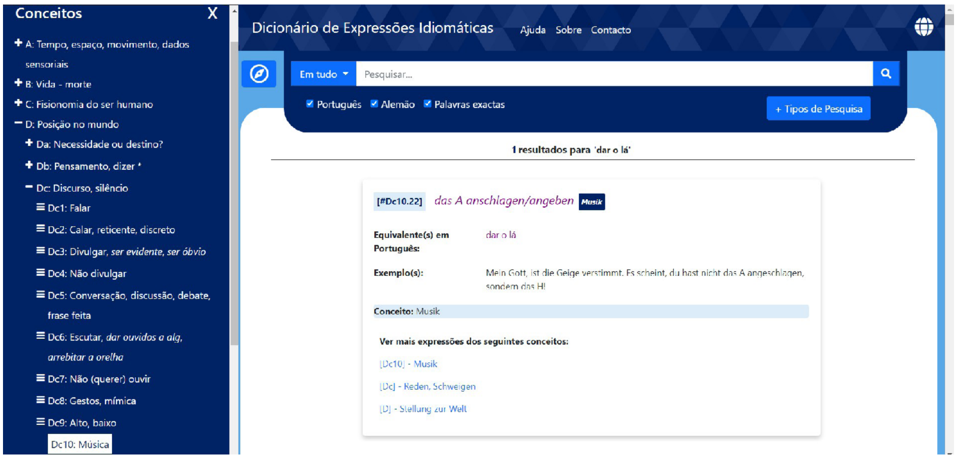
If the user can express the need for information about a language problem, then the interface can show the results according to the selected search methods, which are driven by the user's lexicographical situation.

Last but not least, the interface design was created considering the recent research in interface and usability design for online dictionaries, namely carried out by the Institut für Deutsche Sprache (IDS), Wolfer et al. (2018a, 2018b), Fuertes-Olivera (2016), Bergenholtz et al (2015), Lew & de Schryver (2014), Müller-Spitzer et al (2011, 2014b), and by analysing the online dictionaries mentioned in Section 2, among others.

5.3 Developed Prototype

The current prototype for the search interface (Figure 8) has the following layout design:

- a left side menu to search for idioms by concept, preserving Schemann's original conceptual system, which can be open (compass button) and closed at any time;
- a top menu with the tabs *Ajuda* [Help], *Sobre* [About], *Contacto* [Contact] and a globe button to select the interface language (Figure 8 shows the Portuguese interface[22]);
- a search box, which is always displayed to perform a full-text search at any moment. Note that this search is performed in the entry idioms and in the examples. It is also possible to restrict the search looking for results in *Português* [Portuguese], *Alemão* [German] or searching by *Palavras exactas* [exact words]. It is also possible to restrict the search on specific parts of the dictionary entries: *Em tudo* [all] or only in idioms or concepts and a button '+ *Tipos de Pesquisa'* [+ Search Methods] to focus on the results to assist in writing, idiom comprehension, translation and also advanced search with boolean operators;
- the search results area, showing the dictionaries' entries according to the performed query, which can provide information related to the German idiom, its usage, its Portuguese equivalent(s), the associated concept and synset, as well as suggestions to check for more idioms within the same concept and its higher-order concepts: for example, as shown in Figure 8, there are hyperlinks to search for more idioms in the concepts Dc10 - *Musik* [Music] < Dc - *Reden, Schweigen* [Speaking, Silence] < D - *Stellung zur Welt* [Attitude to the world].


Conceitos X
+ A: Tempo, espaço, movimento, dados sensoriais
+ B: Vida - morte
+ C: Fisionomia do ser humano
− D: Posição no mundo
+ Da: Necessidade ou destino?
+ Db: Pensamento, dizer *
− Dc: Discurso, silêncio
≡ Dc1: Falar
≡ Dc2: Calar, reticente, discreto
≡ Dc3: Divulgar, *ser evidente, ser óbvio*
≡ Dc4: Não divulgar
≡ Dc5: Conversação, discussão, debate, frase feita
≡ Dc6: Escutar, *dar ouvidos a alg, arrebitar a orelha*
≡ Dc7: Não (querer) ouvir
≡ Dc8: Gestos, mímica
≡ Dc9: Alto, baixo
Dc10: Música
Dicionário de Expressões Idiomáticas
Ajuda
Sobre
Contacto
Em tudo
Pesquisar...
☑ Português
☑ Alemão
☑ Palavras exactas
+ Tipos de Pesquisa

1 resultados para 'dar o lá'

[#Dc10.22] *das A anschlagen/angeben* Musik

Equivalente(s) em Português: dar o lá

Exemplo(s): Mein Gott, ist die Geige verstimmt. Es scheint, du hast nicht das A angeschlagen, sondern das H!

Conceito: Musik

Ver mais expressões dos seguintes conceitos:

[Dc10] - Musik

[Dc] - Reden, Schweigen

[D] - Stellung zur Welt

Figure 8 - Search result for '*dar o lá*' [*to hit A*], showing 1 result.

6 Conclusions and Future Work

There are more offers in the market of lexical resources prepared for the decoding task, than for encoding. Studies reveal that "the use of a dictionary for assistance with writing is very high (...) [and] many lexicographers recognise users need dictionaries to look for a word that has escaped their memory although they remember the concept" (Sierra, 2000). For this kind of task, the *Synonym Dictionary*'s onomasiological structure is very useful, namely given its focus on idioms. However, searching an analogical ideological or synonym dictionary can be difficult for some users who might not know how to search in these dictionaries. Thus, a good solution is to make them available in digital format, with their content correctly annotated using XML/TEI. TEI-encoding adds a layer of meaning with more metadata and better data structure, preserving the original content but also leaving space for improvement. If the dictionaries' content is well-

[22] The current prototype is only available with its interface in Portuguese.

annotated with appropriate TEI elements and is well-formed in a solid structure, then the search results will be of more relevance to the user as long as these are presented clearly to assist in idiom learning and/or translation.

The interface design of online dictionaries should break the 'codex layout' that the users know, displaying information in more attractive and interactive ways, to grab the users' interest and attention and give them what they need. This includes the scenarios where not even the users know exactly what they need or are looking for and the search interface should offer suggestions to try to meet what they have in mind. Also, it is important to guide the user throughout the interface whenever it is possible, by offering several search methods, so he/she is not so dependent on his/her searching skills. These search methods are built around the users' needs and will become particularly useful if one identifies his/her needs of information, for instance, assistance in writing a text, understanding an idiom by its synonyms and/or contextualized examples and translating idioms.

We hope that our project provides more direct access to the dictionaries' content and to do so, we will continue to improve it by: i) working on the cross-reference system, ii) adding further annotation within the idioms, iii) adding more keywords related to the top-level and mid-level concepts, so the user will have better chances in finding what he/she needs, and iv) adding examples where there are none. Regarding the examples for Portuguese idioms, these will also be included from *Schemann's Portuguese-German Idiomatic Dictionary.*

Furthermore, a thorough test with the targeted users must be carried out, which may lead to new design specifications.

We believe that our model can later include other languages by encoding four more bilingual Idiomatic Dictionaries of the same author, namely in English, Spanish, French and Italian.

Acknowledgements

This project was funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

References

Bergenholtz, H., Bothma, T., & Gouws, R. (2015). Phases and steps in the access to data in information tools. *Lexikos, 25*(1). https://doi.org/10.5788/25-1-1289

Dias, I. (2010). *Sinonímia - campo semântico - contexto - texto: uma análise sinonímia com particular relevância para as expressões idiomáticas: estudo sistemático e contrastivo*. PhD Thesis. Braga: Universidade do Minho. http://hdl.handle.net/1822/11263

Fuertes-Olivera, P. & Tarp, S. (2014). *Theory and Practice of specialised online dictionaries*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110349023

Fuertes-Olivera, P. (2016). A Cambrian explosion in lexicography: Some reflections for designing and constructing specialised online dictionaries, *International Journal of Lexicography*, Volume 29(2), 226-247. https://doi.org/10.1093/ijl/ecv037

Fuertes-Olivera, P., & Bergenholtz, H. (2018). Dictionaries for text production. In P. Fuertes Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 267-283).

Gouws, R. (2010). Outer texts in bilingual dictionaries. *Lexikos, 14* (pp. 67-88). https://doi.org/10.5788/14-0-683

Klosa, A. (2013). 26. The lexicographical process (with special focus on online dictionaries). In R. Gouws, U.

Heid, W. Schweickard & H. Wiegand (Eds.), *Supplementary Volume Dictionaries. An international Encyclopedia of Lexicography* (pp. 517-524). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.517

Leroyer, P. (2018). Dictionaries for text reception. In P. A. Fuertes-Olivera (Ed.), *The Routledge handbook of lexicography* (pp. 250-266).

Lew, R. & de Schryver, G. (2014), Dictionary users in the digital revolution, *International Journal of Lexicography*, Volume 27(4). (pp. 341-359). https://doi.org/10.1093/ijl/ecu011

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41, https://doi.org/10.1145/219717.219748

Moerdijk, F., Tiberius, C., & Niestadt, J.(2008). Accessing the ANW dictionary. In M. Zock, & C-R. Huang (Eds.), 22nd International Conference on Computational Linguistics. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (pp. 18-24). Brighton, UK: One Digital.

Müller-Spitzer, C., Koplenig, A., & Töpel, A. (2011). What makes a good online dictionary? - Empirical insights from an interdisciplinary research project. *Proceedings of eLex 2011*. (pp. 203-208).

Müller-Spitzer, C. (2014a). 11. Textual structures in electronic dictionaries compared with printed dictionaries: A short general survey. In R. Gouws, U. Heid, W. Schweickard & H. Wiegand (Ed.), *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography* (pp. 367-381). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.367

Müller-Spitzer, C. (Ed.). (2014b). *Using online dictionaries*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110341287

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 in action: Improving the encoding of the dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 417-433). Sintra, Portugal.

Simões, A., Salgado, A., Costa, R. & Almeida, J. J. (2019). LeXmart: A smart tool for lexicographers. In I.

Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 453-466). Sintra, Portugal.

Sierra, G. (2000). The onomasiological dictionary: a gap in lexicography. *Proceedings of the 9th Euralex International Congress* (pp. 223-235). Stuttgart.

Tarp, S. (2012). Online dictionaries: today and tomorrow. *Lexicographica, 28*(2012), 253-268. https://doi.org/10.1515/lexi.2012-0013

Tarp, S. (2013). *Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume: Recent Developments with Special Focus on Computational Lexicography.* Gouws, R. H., Heid, U., Schweickard, W. & Wiegand, H. E. (eds.). New York: De Gruyter, Vol. 5.4. pp. 460-468.

Tarp, S. (2014). Dictionaries in the internet era: Innovation or business as usual? (Enrique Alcaraz Memorial Lecture 2014). *Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses*, (27), 233-261. https://doi.org/10.14198/raei.2014.27.13

TEI Consortium, eds. (2021). Chapter 9: Dictionaries. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.4.2, last modified on 9th April. TEI Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (Accessed May 24, 2021).

Wiegand, H. E. (1990). Printed Dictionaries and their Parts as Text. An Overview of More Recent Research as an Introduction. *In Lexicographica, 6,* 1-126.

Wolfer, S., Nied, M., Dias, I., Müller-Spitzer, C. & Domínguez, M. J. (2018a). Combining quantitative and qualitative methods in a study on dictionary use. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek, Proceedings of the XVIII EURALEX International Congress – Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, pp. 101-112.

Wolfer, S., Kosem, I., Lew, R., Müller-Spitzer, C., & Silveira, M. R. (2018b). Web-based exploration of results from a large European survey on dictionary use and culture: ESDexplorer. *Lexikos*, 28, 440-447. https://dx.doi.org/10.5788/28-1-1473

Dictionary References

Schemann, H. (1993). *Deutsche Idiomatik. Die deutschen Redewendungen im Kontext*. Stuttgart/Dresden: Ernst Klett.

Schemann, H., Schemann-Dias, M. L., Amorim-Braun, L., Martins, T., Duque-Gitt, M.J. & Costa, H. (2012). *Idiomatik Deutsch-Portugiesisch* (2nd ed.). Hamburg: Buske.

Schemann, H. (2012). *Synonymwörterbuch der deutschen Redensarten* (2nd ed.). Berlin: De Gruyter

ENGLISH-CHINESE/CHINESE-ENGLISH CORPUS LEXICOGRAPHY IN CHINA: A REVIEW

Anmin Wang
Guangxi University for Nationalities, Nanning, China 510003
anmin.wang@gxun.edu.cn

Abstract

Corpora have started to play an increasingly important role in dictionary-making in the global context. Against this background, the paper reviews the state-of-art development in English-Chinese/Chinese-English (E-C/C-E) corpus lexicography in China. The E-C/C-E corpus lexicography is reviewed within Li's (2015) framework, from three aspects, corpus- based E-C/C-E dictionary-making, research on corpus-based E-C/C-E dictionaries, and corpus- building for making such dictionaries. As for corpus-based dictionary-making, even if there are already fully corpus-based E-C/C-E dictionary projects, either accomplished or ongoing, their total is still small. In the meantime, a corpus-based general C-E dictionary is still something to come. Using corpora for making specialized E-C/C-E dictionaries would make welcome contributions as well. The research on corpus-based E-C/C-E dictionaries needs to be strengthened. A lot of the research only looks into the role of corpora in E-C/C-E dictionary- making generally, reviews the corpus-based E-C/C-E dictionaries, or investigates how such corpora can help with dictionary-making on the micro-level. Thus researchers can endeavor to investigate how corpora can inspire E-C/C-E dictionary-making in other aspects. Building corpora for making E-C/C-E dictionaries needs to be enhanced, particularly those for making C-E encoding dictionaries and specialized E-C/C-E ones. Overall, there is still much space for the further development of E-C/C-E corpus lexicography in China.

Keywords: corpus lexicography, corpus-based dictionary-making, studies on corpus-based dictionaries, corpus-building for dictionary-making; E-C/C-E dictionaries

1. Introduction

Since the emergence of the monumental *Collins CoBuild English Language Dictionary* in 1987, corpora have started to play an ever increasingly important as well as indispensable role in making a dictionary globally. They have been proved helpful in selecting a headword, labeling its part of speech, defining it, providing illustrative examples, highlighting typical collocations, differentiating synonyms, offering information on its usage frequency, ordering the senses, etc. Other well-known English learner's dictionaries, like *Oxford Advanced Learner's Dictionary*, immediately follow suit. It is no exaggeration to say that being corpus-based has become one of the defining features as well as selling point for current dictionaries globally. Being corpus- based in this paper means that corpora are used as essential tools to play the various roles mentioned above in dictionary-making.

This paper reviews the current status of English-Chinese/Chinese-English (henceforth shortened as E-C/C-E) corpus lexicography in China within the framework of Li (2015: 13-14) against this background. Arguably, with a corpus-based dictionary of Chinese still something to come, the status of E-C/C-E corpus lexicography in China can reveal the overall picture of corpus lexicography in China, since English is the dominant foreign language for the learners in China. The three sections below focus on corpus-based E-C/C-E dictionary-making, theoretical lexicographical explorations of it, as well as corpus-building for making E-C/C-E dictionaries respectively.

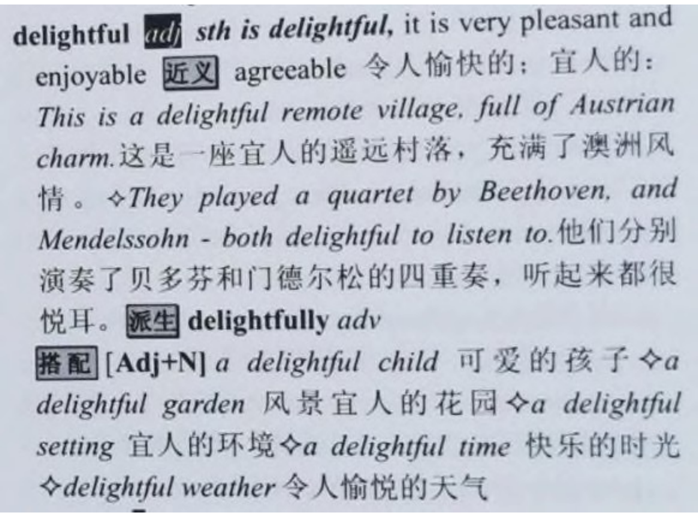
2. The Status of Corpus-based E-C/C-E-Dictionary-making

Building a specialized corpus for making dictionaries, including E-C/C-E ones, involves tremendous work. And thus it is not surprising at all that only a few corpus-based C-E or E-C dictionary projects have been undertaken to date, with some having been accomplished while others still ongoing, to be discussed chronically.

New Age English-Chinese Dictionary(Zhang, 2004) is the first corpus-based E-C dictionary project, which boasts about the inclusion of 150, 000 words and phrases. The CONULEXID, the corpus on which the dictionary is based, has provided tremendous help in selecting headwords, defining them, furnishing various examples, and arranging senses following the order of frequency (e.g.: Zhao, 2005; Mao & Mao, 2005), together with other features like differentiating synonyms, providing quality translation for the examples (Chen, 2005: 53; Chen, 2005). Highly acclaimed as the first large corpus-based learner's dictionary ever compiled in China, it has won various distinguished national or provincial prizes.

Roughly a decade later, *New Century English-Chinese Dictionary* (Hu, 2016) was published, the joint work between two renowned publishers in China and Britain, Foreign Language Teaching and Research Press and HarperCollins UK. The dictionary, compiled by using DPS (Dictionary Production System) developed by Ingénierie Diffusion Multimédia in France, is based on the giant but ever-expanding 4.5-billion-word Collins Corpus. The dictionary of over 250, 000 entries is claimed to be the largest comprehensive E-C dictionary in China. It is "customized" to meet the needs of Chinese users, with headword selection, cultural and encyclopedic information, etc., designed by Chinese lexicographers, and the authentic language data selected out of the Collin Corpus (Xie, 2017: 37).

In addition, another corpus-based E-C dictionary is being compiled by the Center for Lexicographical Studies of Guangdong University of Foreign Studies. (The information is from face-to-face and online communication with one of the major compilers of the dictionary.) The E-C dictionary targets intermediate English learners, with the examples selected from existing corpora like BNC. One of the most innovative features of the dictionary is that a headword is defined by revealing its typical construction first, followed by its definition, as shown in Figure 1 below. Defining the italicized construction in bold may help dictionary users acquire the construction and put it into active use.



delightful adj ***sth is delightful,*** it is very pleasant and enjoyable 近义 agreeable 令人愉快的；宜人的：*This is a delightful remote village. full of Austrian charm.*这是一座宜人的遥远村落，充满了澳洲风情。✧*They played a quartet by Beethoven. and Mendelssohn - both delightful to listen to.*他们分别演奏了贝多芬和门德尔松的四重奏，听起来都很悦耳。派生 **delightfully** *adv*

搭配 **[Adj+N]** *a delightful child* 可爱的孩子 ✧*a delightful garden* 风景宜人的花园 ✧*a delightful setting* 宜人的环境 ✧*a delightful time* 快乐的时光 ✧*delightful weather* 令人愉悦的天气

Figure 1. A sample entry from the dictionary being compiled

One C-E dictionary targeting native Chinese users, based on Parallel English-Chinese Corpus (PECC), is still under way. The headword selection, part-of-speech labeling, digging out the appropriate translation equivalents, furnishing quality examples, particularly those featuring contextualized translation of a headword, and so on, make up the major contributions of PECC to the dictionary. However, due to the lack of compiling staff, entries under several letters are left unfinished (According to the face-to-face communication with the chief editor of the dictionary). Since large corpus-based C-E dictionary is still missing in the bilingual lexicographical field in China, the dictionary is really something to look forward to.

In addition, there are still several specialized corpus-based C-E dictionaries worthy of being discussed. The dictionary projects all focus on culture-loaded terms, though in different ways. The first type of such dictionaries are derived from the parallel corpora of a classic Chinese novel and its different English translations. Liu (2012) compiled a dictionary of 538 Chinese idiomatic expressions based on the parallel corpus of *Sanguo Yanyi* and its two English translations. Two years later, he compiled a C-E translation dictionary based on the parallel C- E corpus of *Shui Hu Zhuan* and its four English versions, with about 2800 entries (Liu 2014). The Chinese headwords are divided into 29 sub-types, including *junshi* (military affairs), *wuqi* (weapons), *fushi* (dress and personal adornment). Liu (2015) developed a parallel corpus based on the classic novel *Rulin Waishi* and its English version and compiled a dictionary of about 1000 headwords of 10 types, including *fal*ü (law), *xiyu* (idiomatic expressions), *yiyao* (medicine), etc. Liu Zequan has adopted the same strategy and compiled *Honglou Meng Hanying Wenhua Da Cidian* on the basis of Honglou Meng and its four widely acclaimed English translations. The dictionary has over 3800 entries, covering eight major categories like *shuyu* (idiomatic expressions), *yinshi*/*fushi*/*yiyao*/*qiyong* (food and drink/dress and personal adornment/medicine/household utensils), and *chengwei yu* (address forms). It will come out soon. The four dictionaries above are not dictionaries proper, but specialized ones offering a list of culture-loaded items in Chinese classic novels and their different translations. Nevertheless, such dictionaries can help users by inspiring them to render the same or a similar items properly into English in context. Another similar dictionary project is being carried out by the Center for Lexicographical Studies of Guangdong University of Foreign Studies with about 4000 culture-loaded Chinese words or expressions, based on the self-constructed 100- millon-word corpus, to appear in the near future.

Overall, the corpus-based E-C/C-E dictionaries are still quite few. Compared with E-C ones, C-E ones are in more urgent need. Even if great progress has been made in making corpus-based E-C dictionaries, there is space for further improvement. Since various learner corpora of different sizes are already available in China, it would benefit Chinese English learners if typical errors committed by them, overuse and underuse of particular word, construction or grammatical phenomenon, etc., are systematically integrated into E-C learner's dictionaries to compile. In the meantime, a corpus-based encoding E-C dictionary targeting foreign learners of Chinese would make a welcome contribution as well. In addition, corpus-based C-E dictionaries for meeting Chinese users' encoding needs are urgently demanded. Meanwhile, apart from improving those corpus-based E-C/C-E dictionaries of culture-loaded items, other corpus-based specialized bilingual dictionaries focusing on the lexical items from various fields or disciplines are also greatly needed. All in all, there is still much room for corpus-based E-C/C-E dictionaries to grow in China.

3. The Status of the Studies on Corpus-based E-C/C-E Dictionaries

Since very few E-C/C-E dictionaries have been compiled by using corpora as a tool so far, it is quite expected that studies on corpus-based E-C/C-E dictionaries are mostly theory-oriented rather than dictionary-making-based. Several points are worthy of being noted as follows. Some studies just focus on the actual or potential impact of corpora in making a general or specialized E-C/C-E dictionary (e.g.: Ye & Zhang, 1997; Li, 2006b; Wang, 2007; Rundell 2009a, 2009b). Meanwhile, there are also researchers exploring how a specialized corpus can contribute to making a specialized dictionary generally, like terminology dictionary (Wang & Wang, 2009; Wang & Wang 2010), E-C navigation dictionary (Wang & Sui, 2015), learner's business English dictionary (e.g.: Hu & He, 2013) and a medical dictionary (Xu, 2017). Cheng, et al. (2019) makes use of a small specialized parallel corpus of medicine to produce a miniature multilingual medical dictionary. The impact of corpora usually involves selecting headwords, offering examples, providing information on frequency, rendering help in revising a dictionary, providing authentic examples, etc. Thus, the conclusions the researchers above have derived are rather similar.

Quite a number of corpus-based dictionary studies are simply the general reviews of a corpus-based monolingual English dictionary, or an E-C/C-E one. The functions of corpora in making English learner's dictionaries are examined generally (e.g.: Yuan, 2000: 80-81; Chen, 2005; Wang & Tian 2013), including *Oxford Advanced Learner's Dictionary* (Yang & Li, 2003) and the Collins dictionaries (Yang & Zhang, 2001; Luo & Cao, 2003), or a specific section like "Wordfinder" (Zheng & Xia 2018) or "Express youself" (Liu & Xia, 2018). Wang (2001) argues Brown Corpus for compiling *A Dictionary of English Collocation* is too small for the purpose and its definition of collocations is too loose. The review of *New Age English-Chinese Dictionary is* capable of serving as a typical example for evaluating a corpus-based bilingual dictionary. (e.g.: Mao & Mao 2004; Yin, 2005; Zhao, 2007). Those studies also mostly focus on the above-mentioned general functions of corpora in making dictionaries.

Some studies go one step further and investigate how a particular corpus can help with enhancing the quality of the micro-structure of a dictionary in various ways. It can render help in determining the part of speech of an English headword (Liu, 2002). Using corpora to help label a Chinese headword with more than one part of speech can be conducive to the provision of more accurate equivalents, thus helping improve the quality of C-E dictionaries (e.g.: Zhang, 2011; Fang, et al., 2011; Ye, 2014; He, 2015; Gong, 2017; Deng, 2018; Wang & Zhang, 2018; Cai & Liu, 2019; Yu, 2020). A parallel E-C corpus can also help dig out and furnish better equivalents or more and better contextualized translations of a headword (e.g.: Li, 2009; Wu & Wang 2012; Li & Zhang, 2016; Ma & Wu, 2016; Ma, 2017), and also collocational information on equivalents (Xia, 2012). In addition, a corpus can be conducive to investigating gender bias in a dictionary (Kuang, 2009), differentiating synonyms (Zhu & Ma, 2010; Ma, 2017), labeling registers (Wang, 2008), and so on.

Few researchers have looked into the corpus-based dictionary-making systems, including the ones for making bilingual E-C/C-E dictionaries. Li (2006a, 2006c, 2006d) introduces the C-E dictionary-making system developed by himself, which has been used for making an encoding C-E dictionary. Chang (2006) elaborates the dictionary-making system designed by Peiking University. A similar system has been developed by Guangdong University of Foreign Studies (Liu, 2006). There are also dictionaries typesetting system and dictionary quality assurance one developed by different institutions available as well. Some researchers try to build corpus-based desk dictionaries based on Lucene (e.g.: Tan & Jiang, 2010), which make looking up a word both in the headword list and in the microstructure of the headword possible instantly. Such desk ones can also present a larger context for a headword than the regular examples in an average dictionary.

Different from the majority of the researchers who hail the benefits of corpora in making a dictionary, a few have paid attention to the potential negative effects in this regard. The size and representativeness of corpus data can impact furnishing a good example (e.g.: Li, 2006, 2015; Wang & Ma, 2003: 26), selecting a potential headword based an its frequency (e.g.: Li, 2006b), defining a word accurately (e.g.: Wang & Ma, 2003: 24-25; Tarp, 2016), and so on. The synchronic nature of corpus data may also lead to problems like imbalance in providing encyclopedic information in a dictionary and errors in definition.

With regard to the studies on corpus-based bilingual dictionary-making, several gaps can be identified immediately. To start with, the implication of corpora for dictionary-making beyond the microstructure needs to be strengthened. Topics like how corpora can help with headword selection, differentiation of synonyms, providing pragmatic, cultural information, etc., can be explored further. In addition, researchers can also dive into how the information derived from corpora can be integrated into an E-C or C-E dictionary, like special difficulties of Chinese English learners, their overuse or underuse of a particular word or construction, their typical errors, etc. Exploring how CEFR-China can be integrated into the corpus-based C-E/E-C dictionaries should be encouraged, including sense arrangement, labeling the frequency of a particular lexical item, etc. Such endeavors can help make the corpus-based dictionaries really tailored to their needs. Thirdly, more research on how to make use of specialized corpora to compile a specialized bilingual dictionary would make a welcome contribution, since very few specialized bilingual dictionaries have been corpus-based. Fourthly, exploring the various large existing corpora, like the huge parallel E-C corpus constructed by Prof. Wang Kefei, to help make quality C-E dictionaries, can be encouraged. Fifthly, since large online corpora have the potential to serve as a dictionary for users, researchers can look into how to design an interface for the corpora, including parallel E-C ones, to make them accessible to users as dictionaries. How to utilize a parallel corpus to generate a dictionary automatically also deserves researchers' great attention. Finally, the research on the dictionary-making systems can be further improved as well, to make them work well for making E-C/C-E bilingual dictionaries. All in all, studies on corpus-based E-C or C-E dictionaries still have quite some space to improve.

4. Corpus-building for Dictionary-making

Against the background that corpora are playing an increasingly important role in dictionary-making, very few corpora have been constructed specifically for dictionary-making in China. To the author's knowledge, only seven different corpora have been specially built for making an E-C/C-E dictionary, at least partly.

The first must-mention is CONULEXID, the first corpus constructed for dictionary- making in China. The 3.5-million-word corpus, with data of different styles, registers, sources, etc. neatly balanced, covers the materials from science, social sciences and others (Zhang, 2001). It was put into use in 1998 for compiling *New Age English Chinese Dictionary*, the first- ever corpus-based large English learner's dictionary in China.

The next dictionary-making-oriented corpus is 20-million-token Parallel English Chinese Corpus (PECC), constructed by Li (2006a, 2006c, 2006d, 2008) for making an encoding C-E dictionary for Chinese users. It is a balanced parallel corpus with materials selected from essays, fictions, etc.， aligned mostly on sentence level. The E-C texts comprise 60% of the corpus data, while the C-E texts 40%. The E-C texts can help ensure the translation equivalents and illustrative examples authentic and native. Meanwhile, one may fail to find the translation equivalents for a Chinese word or expression in the parallel E-C texts, particularly those culture- or politics-loaded ones peculiar to Chinese culture. Then parallel C-E data would play a supplementary but indispensable role in dictionary-making.

Another four parallel C-E corpora were built for making dictionaries of idiomatic or culture-loaded terms. Liu Zequan constructed a sentence-aligned C-E parallel corpus of Hong Lou Meng and its four English translations, with about 2.9 million tokens, which was later used for compiling a C-E dictionary of cultural terms in the novel (Zhang & Liu 2015). Liu (2012, 2014, 2015) constructed three parallel C-E corpora, which are all aligned on the sentence basis, for making C-E dictionaries of idiomatic or culture-loaded Chinese words or expressions. With *Sanguo Yanyi* (Romance of Three Kingdoms) and its two different translations as the basis, Liu constructed a parallel corpus of about 1.76 million tokens. Using the same method, Liu later built another two parallel corpora of *Shuihu Zhuan* and its four English translations, *Rulin Waishi* and its translations, also aligned on the sentence basis. The four corpora, however, are of limited use in dictionary-making because of the size and distribution of data.

In addition, to make a dictionary of Chinese cultural terms, Guangdong University of Foreign Studies constructed a 100-million-token corpus, based on E-C or C-E journal articles, books, authoritative English writings in famous English media, etc., which focus on Chinese culture, language, society, and so on. The dictionary compilation is still under way.

To sum up, the corpus-building for making E-C/C-E dictionaries in China still lacks behind. This is especially true for building large parallel E-C corpus for making encoding C-E dictionaries for Chinese users. Despite the fact that there are quite a number of parallel E-C corpora available for use online (Wang & Huang, 2012), very few of them have been constructed specifically for dictionary-making. For example, the parallel E-C Corpus constructed by Prof. Wang Kefei has around 100 million tokens. However, it is mainly made up of texts from literary translations. As Li (2015) points out, corpora built for dictionary-making needs to meet their own criteria, like the appropriate size, texts covering different registers, being balanced in the text length. Therefore, the distribution of the data in Prof. Wang's corpus needs to be changed or modified before it can be used for making a quality encoding C-E dictionary. In the meantime, the construction of quality English corpus or E-C corpora for producing good E-C or English dictionaries need be planned and carried out as well. Parallel corpora for making E-C/C-E dictionaries of different disciplines or fields are also highly needed. In a word, the corpus-building for E-C/C-E dictionary-making needs to be greatly strengthened.

4. Conclusion

This paper reviews the E-C/C-E corpus lexicography in China within the framework proposed by Li (2015). Despite the fact that corpus-based E-C/C-E lexicography in China has witnessed great progress, there are still various gaps to be bridged. In terms of corpus-based E-C/C-E dictionary-making, C-E dictionaries targeting Chinese users' encoding needs are in greater demand than E-C ones for decoding purposes. Various specialized corpus-based E-C/C-E dictionaries for different disciplines or fields are also greatly needed. When it comes to the research on corpus-based E-C/C-E dictionaries, different topics need to be strengthened as well. Again, looking into how to utilize a parallel E-C corpus to produce an encoding C-E dictionary can be prioritized, among many other ones. Constructing parallel E-C corpora to serve dictionary-making purposes also need to emphasize building those for making an encoding C- E dictionary. Of course, various corpus-based E-C/C-E dictionaries for different disciplines are also in need. In a word, there is still a lot of space for Chinese corpus-based lexicography to develop.

References

Cai, X. Q. & Liu, R. H. (2019). A preliminary study on labeling the part of speech of "yinmi". *Guangdong Canye*, *53*(3), 142-143. http.//doi:CNKI:SUN:GDCY.0.2019-03-089.

Chang, B. B. (2006). Building the corpus-based bilingual dictionary-making platform. *Lexicographical studies*, (3),122-133. http.//doi:10.16134/j.cnki.cn31-1997/g2.2006. 03.017.

Chen, H. W. 2005. New cultural concepts in compiling bilingual dictionaries: A case study of *New Age English-Chinese dictionary*. *Foreign languages and their teaching*, (2), 51-54. http.// doi:CNKI:SUN:WYWJ.0.2005-02-013.

Chen, W. (2005). Seeking for idiomaticity in accuracy and revealing functions in lifelikeness: On the translation of the examples in New Age English-Chinese Dictionary. *Lexicographical Studies*, (2), 33-41. http.//doi:10.16134/j.cnki.cn31-1997/g2.2005. 02.005.

Chen, Y. (2005). The trend in compiling monolingual English learner's dictionaries in the 21st century. *Foreign language and literature studies*, (3), 168-171+211. http.//doi:10. 19716/j.1672-4720.2005.03.006.

Cheng, S. H. et al. (2019). On the translation of Ayurveda vocabulary through quantitative analysis and qualitative analysis within corpus in compiling Sanskrit—Latin—English—Chinese dictionary of Ayurveda. J*ournal of Panzhihua University*, *36*(4),66- 74. htt p:/ /doi :10.13773/j .cnki.51 -1637/z.2019.04. 012.

Deng, J. (2018). An investigation of Labeling the Part of Speech of "Chenggong" in Chinese-English Dictionaries: From the perspective of two—level word class categorization theory. *Journal of Kaifeng Institute of Education*, *38*(7),77-79. http.//doi: CNKI:SUN:KFJY.0.2018-07-037.

Fang, Z. C., Guo, Y. H. & Zhao, L. (2011). Strict Equivalence of Word Category in Chinese- English Dictionary Definition. *Journal of Nanjing College for Population Programme Management*, *27*(4), 75-78. http.//doi:10.14132/j.2095-7963.2011.04.001.

Gong, N. (2017). Labeling the Part of Speech of "tanwu" and its Strategies. *Journal of Chongqing University of Education*, *30*(1), 39-42+51. http.//doi:CNKI:SUN:XQJI. 0.2017-01-008.

He, Y. Q. (2015). Labeling the Part-of-Speech of "zhufu" in Chinese-English Dictionary. *Journal of Henan Institute of Technology (Social Science Edition)*, *30*(1), 70-72+80. http.//doi:10.16203/j.cnki.cn41-1396/c.2015.01.014.

Hu, C. Y. & He, J. N. (2013). A Corpus-based Study of Business English Lexicography for English Learners: A Critical Review of *Oxford Business English Dictionary for Learners of English*. *Journal of Guangdong University of Foreign Studies*, (6), 38-41+91. http://doi:CNKI:SUN:GDWY.0.2013-06-009.

Hu, Z. L. (2016). *New Century English-Chinese Dictionary*. Beijing: Foreign LanguageTeaching and Research Press.

Kuang, Q. (2009). The Gender Controversy in Dictionaries: A case study of gender marker entries based on corpus. *Journal of Hebei Polytechnic University (Social Science Edition)*, *9*(4), 169-172. http://doi:CNKI:SUN:HLXB.0.2009-04-054.

Li, D. J. (2006a). Use of Parallel Corpus in Bilingual Dictionary-making. *Journal of PLA University of Foreign Languages*, *29*(3): 41-44+64. http.//doi:CNKI:SUN:JFJW. 0.2006-03-007.

Li, D. J. (2006b). Reflections on Applying Corpora to Making Bilingual Dictionaries. *Lexicographical Studies*, (2), 104-109. http.//doi:10.16134/j.cnki.cn31-1997/g2. 2006.02.015.

Li, D. J. (2006c). An Introduction to CpsDict—A Bilingual Dictionary-making System Based on Parallel Corpora. *Foreign Languages Research*, (2), 63-65+68+80. http.//doi:CNKI:SUN:NWYJ.0.2006-02-012.

Li, D. J. (2006d). CpsDict: Bilingual dictionary-making system based on parallel corpus. *Modern Foreign Languages*, (4), 371-381+437. http.//doi:CNKI:SUN:XDWY.0.2006-04-006.

Li, D. J. (2008). The Construction of PECC Was Accomplished. *Foreign Language Research*, (6), 73. http.//doi:CNKI:SUN:NWYJ.0.2008-06-018.

Li, D. J. (2015). *Corpus Lexicography: Theory, method and application*. Nanjing: Yilin Press.

Li, N. N. & Zhang, Z. (2016). A COCA Corpus-Based Analysis of the English Translations of the Term *Shehui Gongde*. *Bilingual Education*, *3*(3), 74-80. http.//doi:10.13953/j.cnki. syjyyj.2016.03.012.

Liu, H., Li, Y. Z. & Zhang, Y. H. (2006). Design and Realization of WEB Dictionary Editing and Automatic Generating System Based on Corpus. *Journal of Shenyang Normal University (Natural Science)*, *24*(3), 306-309. http.//doi:CNKI:SUN:SYSX.0.2006-03 -014.

Liu, H. L. Part -of -speech Tagging in Corpus -based Bilingual Dictionary Compilation. *Journal of North China Institute of Technology*, *3*(3), 63-65. http.//doi:CNKI:SUN: HGXS.0.2002-03-022.

Liu, K. Q. (2012). *Corpus Lexicography and Parallel Corpus-based Compilation of Idiom Translation Dictionary of The Romance of Three Kingdoms*. Kunming: Yunnan University Press.

Liu, K. Q. (2014). *The Translation Dictionary of Shui Hu Zhuan*. Beijing: Central Compilation and Translation Press.

Liu, K. Q. (2015). On the Typical Translation of Lexical Items in *Ru Lin Waishi*: Studies based on a parallel corpus. Beijing: Guangming Daily Press.

Liu, Y. & Xia, L. X. (2018). A Study on “Wordfinders” in *Oxford Advanced Learner’ s Dictionary* (9th edition). *Lexicographical Studies*, (5), 31-39+94. http://doi:10.16134 /j.cnki.cn31-1997/g2.2018.05.006.

Luo, S. M. & Cao, J. W. (2003). A Review of Collins English Dictionary (Centennial edition). *Lexicographical Studies*, (5), 87-91+130. http.//doi:10.16134/j.cnki.cn31-1997 /g2.2003.05.012.

Ma, L. D. (2017a). Using Corpora and other Internet Resources to Choose the English Equivalents of Chinese Neologisms: A case study of the English translations of the Chinese keyword Yirenweiben. *Shandong Foreign Language Teaching*, *38*(6), 84-93. http.//doi:10.16482/j.sdwy37-1026.2017-06-010.

Ma, L. D. (2017b). Gathering Lexical Knowledge from Online English Dictionaries and Other Internet Resources: Case Study of Discovering Features That Distinguish FLAMMABLE from Its Synonyms. *China Terminology*, *19*(5), 64-69. http.//doi: CNKI:SUN:KJSY.0.2017-05-014.

Ma, L. D. & Wu, G. H. (2016). Mining Chinese-English Interlingual Equivalents in the Age of Big Data. *Lexicographical Studies*, (3), 43-55+71+94. http.//doi:10.16134/j.cnki.cn 31-1997/g2.2016.03.005.

Mao, B. B & Mao, Z. M. (2005). On Compiling Contemporary Learner’s Dictionaries: From the perspective of *New Age English-Chinese Dictionary*. *Shanghai Journal of Translators*, (S1), 92-93. http.//doi:CNKI:SUN:SHKF.0.2005-S1-026.

Rundell, M. (2009a). Trans. Xia, L. X. & Zhu, D. S. Taking Corpus Lexicography to the Next Level: Explicit Use of Corpus Data in Dictionaries for Language Learners (I). *Lexicographical Studies*, (3), 71-78. http.//doi:10.16134/j.cnki.cn31-1997/g2.2009.03. 024.

Rundell, M. (2009b) Taking Corpus Lexicography to the Next Level: Explicit Use of Corpus Data in Dictionaries for Language Learners (I). Lexicographical Studies, (4), 81-91. http.//doi:10.16134/j.cnki.cn31-1997/g2.2009.04.006.

Tan, G. D. & Jiang, T. (2010). Design and Implementation of Electronic Dictionary Based on Lucene. *China Medical Education Technology*, *24*(5), 510-513. http.//doi:10.13566/j. cnki.cmet.cn61-1317/g4.2010.05.035.

Tart, S. Trans. Xue, M. (2016). Corpus-driven，Corpus-based or Corpus-assisted Lexicography: The Limited Usefulness of Corpora in Defining Specialized Terms. *Lexicographical Studies*, (4), 1-11+93. http.//doi:10.16134/j.cnki.cn31-1997/g2. 2016.04.001.

Wang, D. H. & Wang, L. Y. Summary of the Study of Terminology in the Foreign Language Circle of China. *Lexicographical Studies*, (2),111-123. htt p:/ /doi :10.16134/j.cnki.cn31 -1997/g2.2010.02.005.

Wang, F. F. On the Corpus for A Collocation Dictionary: The review of a dictionary of English collocations. *Lexicographical Studies*, (1), 99-107. http.//doi:10.16134/j.cnki. cn31-1997/g2.2001.01.014.

Wang, F. F. & Ma, L. M. (2003). The Limitations of Corpus-based Dictionaries. *Lexicographical Studies*, (5), 20-28. http.//doi:10.16134/j.cnki.cn31-1997/g2. 2003.05.003.

Wang, H. H. & Zhang, J. (2018). On the Entry of Self-designation Senses into Chinese English Dictionaries: a Corpus-based Study. *Journal of Longdong University*, *29*(6),

17-20. http.//doi:CNKI:SUN:LDXS.0.2018-06-004. Wang, J. & Sui, G. L. (2015). On Compiling Corpus-based Student English-Chinese Navigation Dictionary. *Shandong Social Sciences*, (S2), 375-377. Http:// doi:10.14112 /j.cnki.37-1053/c.2015.s2.160.

Wang, J. S. & Tian, J. G. (2013). Corpus—based Lexicography and Key Issues Concerned. *Journal of Zhengzhou Institute of Aeronautical Industry Management (Social Science Edition)*, *32*(1), 71-73. http.//doi:10.19327/j.cnki.zuaxb.1009-1750.2013.01.019.

Wang, K. F. & Huang, L. B. (2012). Construction and Application of Parallel Corpora: Issues and Comments. *Technology Enhanced Foreign Language Teaching*, 6, 3-10. http.//doi: CNKI:SUN:WYDH.0.2012-06-002.

Wang, L. Y. & Wang, D. H. (2009). A Study on the Terminological Dictionary based on Terminological Education. *China Terminology*, (6), 35-39. htt p:/ /doi :C NKI:S UN: KJSY.0.2009-06-014.

Wang, T. (2008). Register Labeling in English-Chinese Dictionaries. *Journal of Xiamen University of Technology*, *16*(02),107-112. http.//doi:10.19697/j.cnki.1673-4432.2008 .02.021.

Wang, X. H. (2001). On the Impact of Corpora on Dictionary-making. *Lexicographical Studies*, (04),15-21. http.//doi:10.16134/j.cnki.cn31-1997/g2.2001.04.003.

Wu, X. Y. & Wang, A. M. (2012). Parallel Corpus and C-E Dictionary Compilation. *Translations*, (2),169-176. http.//doi:CNKI:SUN:YILN.0.2012-02-021.

Xia, L. X. (2012). Putting Corpus Data into the Collocation Information in Chinese-English Dictionaries for Chinese Users. *Foreign Language and Literature*, *28*(3):47-50. http.// doi:CNKI:SUN:SCWY.0.2012-03-010.

Xie J. X. (2017). On the coordination between DPS-based compilation and publication of A New Century English-Chinese Dictionary: A case study of the annotation function of Entry Editor. *Lexicographical Studies*, (1), 33-43.

Xu, J. J. (2017). Genre-informed Phraseological Approach to Compiling an EAP Dictionary of Medical English. *Foreign Languages and Their Teaching*, (6), 52-60+146. http.//doi: 10.13458/j.cnki.flatt.004443.

Yang, W. & Zhang, B. R. (2001). Pursuing Practicality and Innovation: On compiling contemporary learner's dictionaries exemplified with the Collins English dictionaries. *Lexicographical Studies*, (3),105-112. http.//doi:10.16134/j.cnki.cn31-1997/g2.2001. 03.017.

Yang, X. J. & Li, S. H. (2003). The Advantages of Corpora in Dictionary-making: On Oxford Advanced Learner's Dictionary (6th ed.). *Foreign Languages and Their Teaching*, (4), 47-51. http.//doi:CNKI:SUN:WYWJ.0.2003-04-011.

Ye,G.&Zhang,B.R.English-ChineseBilingualCorporaandtheCompilationofEnglish-ChineseDictionaries. *Journal of Nanjing University*, (1), 166-172. http.//doi:CNKI:SUN:NJDX.0.1997-01-023.

Ye. Z, X. (2014). Corpus-based Study on Parts of Speech of "jiaoqing"and Inspirations in Chinese-English Dictionary Compiling. Journal of Chongqing University of Technology (Social Science), *28*(11), 137-143. http.//doi:CNKI:SUN:CQGS.0.2014-11-026.

Yin, B. Y. (2005). On New Age English-Chinese Dictionary (Social Science Edition). *Journal of Jiangsu University*, *7*(1). http://doi:10.13317/j.cnki.jdskxb.2005.01.018.

Yu, F. (2020). The Study of Word Class Labelling for "Du-li" —A Corpus-based Case Study. *Journal of Chongqing College of Electronic Engineering*, *29*(5), 85-89. http.//doi:10.13887/j.cnki.jccee.

Yuan, K. L. (2000). New Thoughts in 1990s English Learner's Dictionaries. *Lexicographical Studies*, (1), 80-88.

Zhang, J. H. (2011). A Corpus — based Case Study of POS labeling of Chinese—English Dictionaries. *Journal of Sichuan College of Education*, *27*(7): 69-72. http.//doi:CNKI: SUN:SJXB.0.2011-07-020.

Zhang, S. W. (2001). An Introduction to Textual Corpus in the Corpus System of CONULEXID. In ChinaLex Bilingual Committee (eds.) *The Proceedings of the Fourth Annual Symposium and Conference* (pp. 281-291). Nanjing: Jiangsu Educational Press.

Zhang, Bairan. (2004). *New Age English-Chinese Dictionary*. Beijing: The Commercial Press.

Zhang, D. D. & Liu, Z. Q. (2015). The Values of Corpus-based Chinese-English Translation Dictionary: With Reference to the Hong Lou Meng Chinese-English Culture Dictionary. *Journal of Hebei University (Philosophy and Social Science)*, *40*(6):58-64. http.//doi: CNKI:SUN:HBDS.0.2015-06-009.

Zhao, C. H. (2007). A Review of New Age English-Chinese Dictionary. In Fujian Foreign Languages Association. *The Conference Proceedings of Fujian Foreign Languages Association & The Fourth Symposium on Foreign Language Teaching in Eastern China*, pp. 56-60.

Zheng, F. & Xia, L. X. (2018). A Study on "Wordfinders"in Oxford Advanced Learner's Dictionary (9th edition). *Lexicographical Studies*, (5), 22-30+93. http://doi:10.16134 /j.cnki.cn31-1997/g2.2018.05.005.

Zhu, W. H. & Ma, L. D. (2010). Improving the Column of Synonym Differentiation in English-Chinese Learner's Dictionaries: Based on corpus data and their application. *Lexicographical Studies*, (6), 78-87. http.//doi:10.16134/j.cnki.cn31-1997/g2.2010. 06.025.

ADAPTING WORD SKETCHES FOR SPECIALIZED KNOWLEDGE EXTRACTION

Antonio San Martín, Catherine Trekker
University of Quebec in Trois-Rivières, Canada
antonio.san.martin.pizarro@uqtr.ca; catherine.trekker-seguin@uqtr.ca

Abstract

Word sketches (WSs) in Sketch Engine have become a basic tool in terminology work. They bring out patterns of term behavior that would be too time-consuming to identify manually. Most default English WS columns in Sketch Engine extract words with a frequent syntactic relationship with the search word in a corpus (e.g., the nouns usually functioning as the subject of a given verb). The usefulness of the default syntactic WSs for collocational analysis is evident, but their contribution to specialized knowledge extraction is less straightforward.

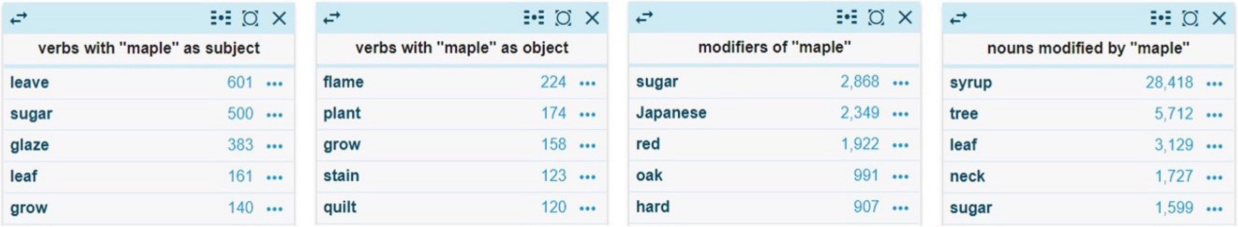
This paper presents a work in progress consisting of adapting the default WSs for specialized knowledge extraction. In previous work, we developed the contextonymic WS, and semantic WSs, which specifically target specialized knowledge extraction. This paper explores two changes to the default WSs. The first change enables WSs to extract nouns functioning as subject and object in the same sentence: (e.g., fertilizer>yield: *fertilizer increases yield*; *fertilizer improves yield*), which usually corresponds to an agent-patient relation. The other change concerns the extraction of the adjectives that modify a noun. This involves modifications to two of the existing WS columns that extract adjectives and the addition of a new type.

We evaluated the precision of these adaptations in a specialized corpus of English texts on Agronomy. Additionally, we compared their output with terminological definitions of a set of terms to assess their usefulness for specialized knowledge extraction. The results indicate that WS columns of nouns functioning as subject and object in the same sentence are sufficiently accurate and potentially useful for specialized knowledge extraction. However, the results for the adjectival WS columns are inconclusive.

Keywords word sketches, corpus analysis, specialized knowledge extraction, Sketch Engine

1 Introduction

One of the most useful features of Sketch Engine (https://www.sketchengine.eu/) (Kilgarriff et al., 2014) is the generation of word sketches (WSs), which have become a basic tool in terminology work. A WS is a one-page summary of a search word's most common usage patterns in a given corpus. It lists the words that are syntactically related to the search word in the corpus and includes a link to the corresponding concordances. Some examples of WS columns are the verbs having the search word as subject or object, the modifiers of the search word, or the words that the search word modifies (Figure 1).



verbs with "maple" as subject	
leave	601
sugar	500
glaze	383
leaf	161
grow	140

verbs with "maple" as object	
flame	224
plant	174
grow	158
stain	123
quilt	120

modifiers of "maple"	
sugar	2,868
Japanese	2,349
red	1,922
oak	991
hard	907

nouns modified by "maple"	
syrup	28,418
tree	5,712
leaf	3,129
neck	1,727
sugar	1,599

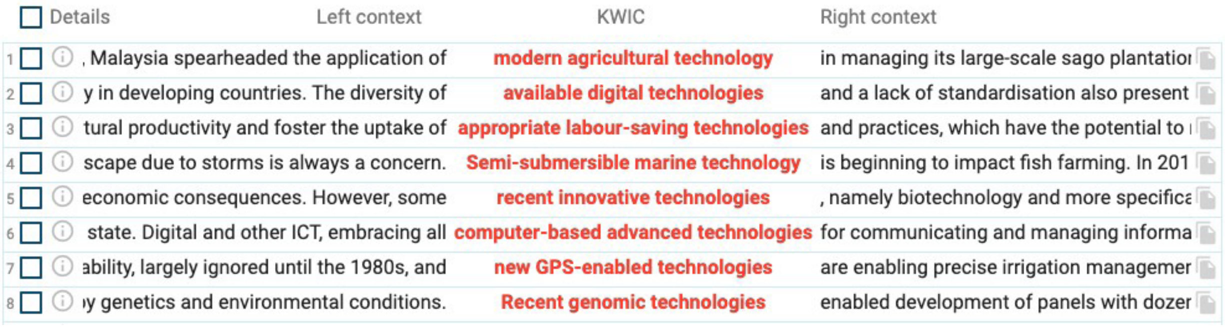
Figure 1. Default WS columns of *maple* in enTenTen18 corpus

Since these word behavior patterns are too time-consuming to identify manually, WSs significantly facilitate collocational analysis. However, corpus analysis is not only useful for extracting linguistic information but also for conceptual knowledge. This is especially true when it comes to elaborating terminological definitions (in which the conceptual content that terms convey is described) or build conceptual networks (in which concepts are interconnected through conceptual relations). For these tasks, the usefulness of the WSs that Sketch Engine generates by default is less straightforward.

In previous work, we proposed new types of WS specifically developed for specialized knowledge extraction: (i) contextonymic WS (see 1.2.1); (ii) semantic WSs (see 1.2.2). This paper presents a work in progress consisting of adapting the default English WS for the same purpose. More specifically, we explore the creation of two WS columns that extract the relation between the subject and the object of the same sentence and the grouping of different columns that extract the adjectives that qualify a noun. This adapted version of the default WS would eventually become part of a single WS that specifically targets specialized knowledge extraction along with the contextonymic WS and the semantic WS.

The rest of the article is organized as follows. In the remainder of this section, we will explain how WSs are generated, and we describe our previous work on creating WSs for specialized knowledge extraction. Section 2 will focus on how the new WS columns were developed and evaluated. Section 3 presents the evaluation results. Finally, in Section 4, we analyze the results and draw some conclusions.

1.1 Word sketch generation



□ Details	Left context	KWIC	Right context
1 □ ⓘ	, Malaysia spearheaded the application of	modern agricultural technology	in managing its large-scale sago plantatioı
2 □ ⓘ	y in developing countries. The diversity of	available digital technologies	and a lack of standardisation also present
3 □ ⓘ	tural productivity and foster the uptake of	appropriate labour-saving technologies	and practices, which have the potential to ı
4 □ ⓘ	scape due to storms is always a concern.	Semi-submersible marine technology	is beginning to impact fish farming. In 201
5 □ ⓘ	economic consequences. However, some	recent innovative technologies	, namely biotechnology and more specifica
6 □ ⓘ	state. Digital and other ICT, embracing all	computer-based advanced technologies	for communicating and managing informa
7 □ ⓘ	ability, largely ignored until the 1980s, and	new GPS-enabled technologies	are enabling precise irrigation managemer
8 □ ⓘ	ıy genetics and environmental conditions.	Recent genomic technologies	enabled development of panels with dozer

Figure 2. Concordances illustrating the rule "[tag="J.*"]{2} [lemma="technology"]"

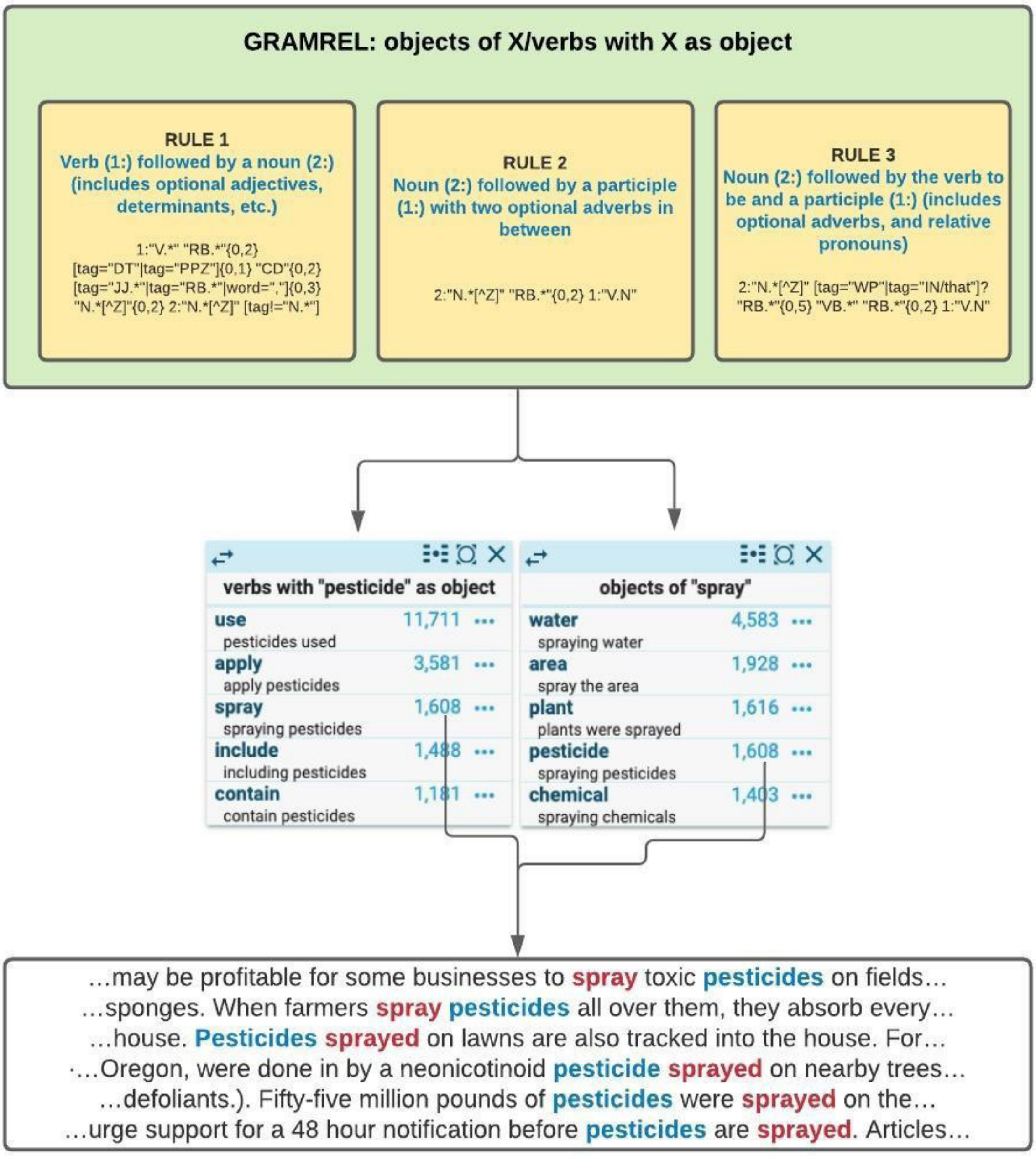
For WS generation, the rules intended to capture the same syntactic relation are grouped into a gramrel (for "grammatical relation"). For instance, to identify the relation between verbs and their objects, the gramrel "objects of X/verbs with X as object" (included in the default sketch grammar) is composed of three rules (Figure 3)

The set of gramrels that produce a WS constitutes a sketch grammar (Figure 4). For example, the default English sketch grammar contains 40 rules organized into 25 gramrels. The number of WS columns can be greater than the number of gramrels because a dual gramrel produces two columns (e.g., "objects of X/ verbs with X as object", which results in one column for the verbs, and another for the nouns).

Since sketch grammars are text files containing CQL rules grouped in gramrels, it is possible to modify or expand them by integrating new rules or adapting or deleting existing ones. Users can compile their own corpora with the sketch grammar of their choice in Sketch Engine. This allows the creation of sketch grammars adapted to different corpus needs.

1.2 Sketch grammars for specialized knowledge extraction

The default sketch grammar in Sketch Engine is mainly based on syntactic co-occurrence. In other words, it lists words that appear in the same context as the search word and which maintain a syntactic relationship with it (Evert, 2009, p. 1222). This type of co-occurrence is of great importance for collocational analysis, and WSs were designed with this end in mind (Kilgarriff & Tugwell, 2001). The usefulness of syntactic co-occurrence for specialized knowledge extraction is less straightforward because the relevance of syntactic relations for conceptual analysis varies. For instance, the WS listing the modifiers of a noun may include, among others, adjectives indicating entrenched hyponyms (e.g., "<u>urban</u> farmer", "<u>rural</u> farmer"), possible features (e.g., "<u>risk-averse</u> farmer", "<u>successful</u> farmer"), or they may not be of interest for conceptual analysis (e.g., "<u>other</u> farmer", "<u>same</u> farmer"). Therefore, syntactic co-occurrence can be exploited for specialized knowledge extraction. However, the default English sketch grammar needs to be specially adapted for that purpose.



GRAMREL: objects of X/verbs with X as object

RULE 1
Verb (1:) followed by a noun (2:) (includes optional adjectives, determinants, etc.)

```
1:"V.*" "RB.*"{0,2}
[tag="DT"|tag="PPZ"]{0,1} "CD"{0,2}
[tag="JJ.*"|tag="RB.*"|word=","]{0,3}
"N.*[^Z]"{0,2} 2:"N.*[^Z]" [tag!="N.*"]
```

RULE 2
Noun (2:) followed by a participle (1:) with two optional adverbs in between

```
2:"N.*[^Z]" "RB.*"{0,2} 1:"V.N"
```

RULE 3
Noun (2:) followed by the verb to be and a participle (1:) (includes optional adverbs, and relative pronouns)

```
2:"N.*[^Z]" [tag="WP"|tag="IN/that"]?
"RB.*"{0,5} "VB.*" "RB.*"{0,2} 1:"V.N"
```

verbs with "pesticide" as object	
use pesticides used	11,711 •••
apply apply pesticides	3,581 •••
spray spraying pesticides	1,608 •••
include including pesticides	1,488 •••
contain contain pesticides	1,181 •••

objects of "spray"	
water spraying water	4,583 •••
area spray the area	1,928 •••
plant plants were sprayed	1,616 •••
pesticide spraying pesticides	1,608 •••
chemical spraying chemicals	1,403 •••

…may be profitable for some businesses to **spray** toxic **pesticides** on fields…
…sponges. When farmers **spray pesticides** all over them, they absorb every…
…house. **Pesticides sprayed** on lawns are also tracked into the house. For…
·…Oregon, were done in by a neonicotinoid **pesticide sprayed** on nearby trees…
…defoliants.). Fifty-five million pounds of **pesticides** were **sprayed** on the…
…urge support for a 48 hour notification before **pesticides** are **sprayed**. Articles…

Figure 3. "Objects of X/verbs with X as object" gramrel, its resulting WS columns and concordances from enTenTen18 corpus

Syntactic co-occurrence lies halfway along a continuum, with surface co-occurrence (the less constraining kind) at one end and semantic co-occurrence (the most constraining kind) at the other end. Surface co-occurrence, on which the contextonymic sketch grammar (see 1.2.1) is based, occurs when two words appear in the same context without the need of any syntactic or semantic relationship (Evert, 2009, p. 1215). For instance, in "Because glyphosate is systemic, excess residue levels can persist…", *glyphosate* and *residue* would be surface co-occurrents (or *contextonyms) (as well as glyphosate and because, is, systemic, etc.), even if they do not establish* a direct syntactic or semantic relation. As for semantic co-occurrence, two words are said to co- occur if a semantic relationship is established between them in a given context (e.g., hyponymy, meronymy, cause, etc.). For instance, in "Glyphosate is the only herbicide that kills…" *glyphosate* and *herbicide* are semantic co-occurrents because there is a hyponymic relation between them in that context.

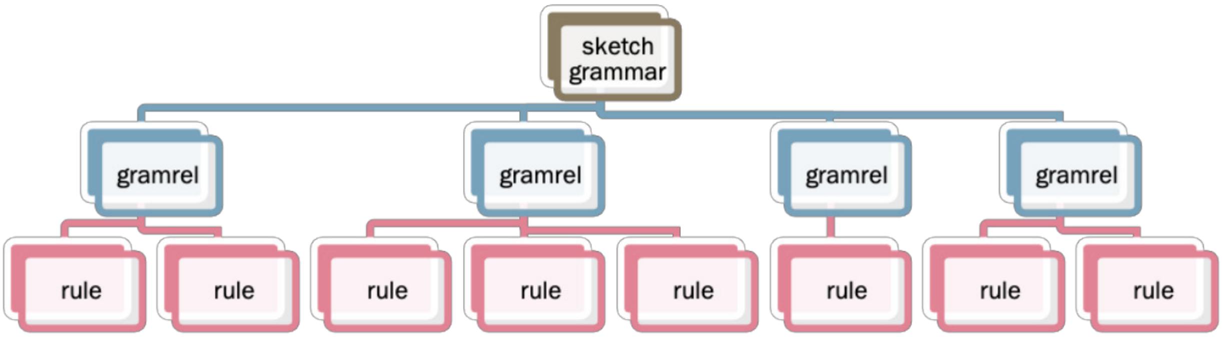


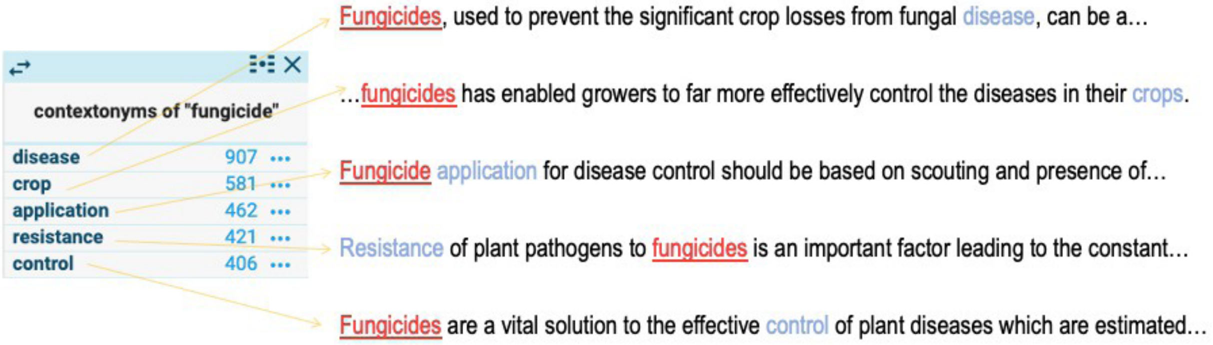
Figure 4. Example of the structure of a sketch grammar

The boundary between these three types of co-occurrence is fuzzy. Surface co-occurrence is generally based on a window of tokens. In contrast, syntactic and semantic co-occurrences are detected by patterns. Although the proposed adaptation of the default WSs in this paper is mostly based on syntactic co-occurrence, some semantic components are also introduced. Before describing the proposed adaptations to the default grammar, we briefly present the contextonymic WS and the semantic WSs since the proposed modifications complement both.

1.2.1 *Contextonymic WS*

Extracting the contextonyms of a term can help to determine its semantic features (San Martín, in press). The contextonymic WS was developed for extracting specialized knowledge for definition writing (San Martín, 2016). Contextonym extraction can be based on various parameters (window span, exclusion of certain parts of speech etc.). The current version of the contextonymic sketch grammar contains one gramrel that defines the contextonym of a word as any verb, noun, or adjective before or after the search word with zero to 44 words between them beyond sentence or paragraph limits. It also excludes certain very common lemmas (e.g., *be, have,* etc.) that do not convey significant semantic features of the search word.

The contextonymic sketch grammar is useful for specialized knowledge extraction because it provides terms that are closely related to the search word, which are sometimes not captured by other WS. For example, by consulting the contextonyms of *fungicide* in our corpus, it is possible to deduce that important semantic features of the term are that fungicide application allows the control of certain diseases in crops, but some pathogens can develop resistance to them. Figure 5 reproduces concordance lines that illustrate the relation of fungicides and its first five contextonyms.



contextonyms of "fungicide"

disease	907
crop	581
application	462
resistance	421
control	406

Fungicides, used to prevent the significant crop losses from fungal disease, can be a...

...fungicides has enabled growers to far more effectively control the diseases in their crops.

Fungicide application for disease control should be based on scouting and presence of...

Resistance of plant pathogens to fungicides is an important factor leading to the constant...

Fungicides are a vital solution to the effective control of plant diseases which are estimated...

Figure 5.Contextonymic WS and concordances of *fungicide* in our Agronomy corpus

With the contextonymic sketch grammar, it is usually necessary to consult the corresponding concordances to discover their relation to the search word. This disadvantage is compensated for by the fact that this WS yields valuable results even in smaller corpora.

1.2.2 *Semantic sketch grammar*

Semantic co-occurrence is based on knowledge patterns, which are lexico-syntactic patterns that match contexts in which a specific semantic relation is conveyed (Meyer, 2001, p. 281). An example of a knowledge pattern is “X and other Y” (e.g., “manure and other fertilizers”), which encodes a hyponymic relation (*manure* is a type of *fertilizer*), or “X contains Y” (e.g., “*fertilizers* contain urea”), which encodes a meronymic relation (*urea* is a part of *fertilizer*).

The EcoLexicon Semantic Sketch Grammar (ESSG) (http://ecolexicon.ugr.es/essg/) (León-Araúz et al., 2016; León-Araúz & San Martín, 2018) encodes knowledge patterns that capture hyponymy, meronymy, cause, function, and location relations in English. There is a French version that at the moment only includes hyponymy (San Martín et al., 2020). Figure 6 shows an example of each of the columns of the ESSG in English extracted from the EcoLexicon corpus (León-Araúz et al., 2018). This corpus is available to any Sketch Engine user and comes compiled with the ESSG.

The ESSG has the advantage of clearly identifying the semantic relationship linking the terms, but the number of results is lower than with other types of WS and requires large corpora to yield useful results.

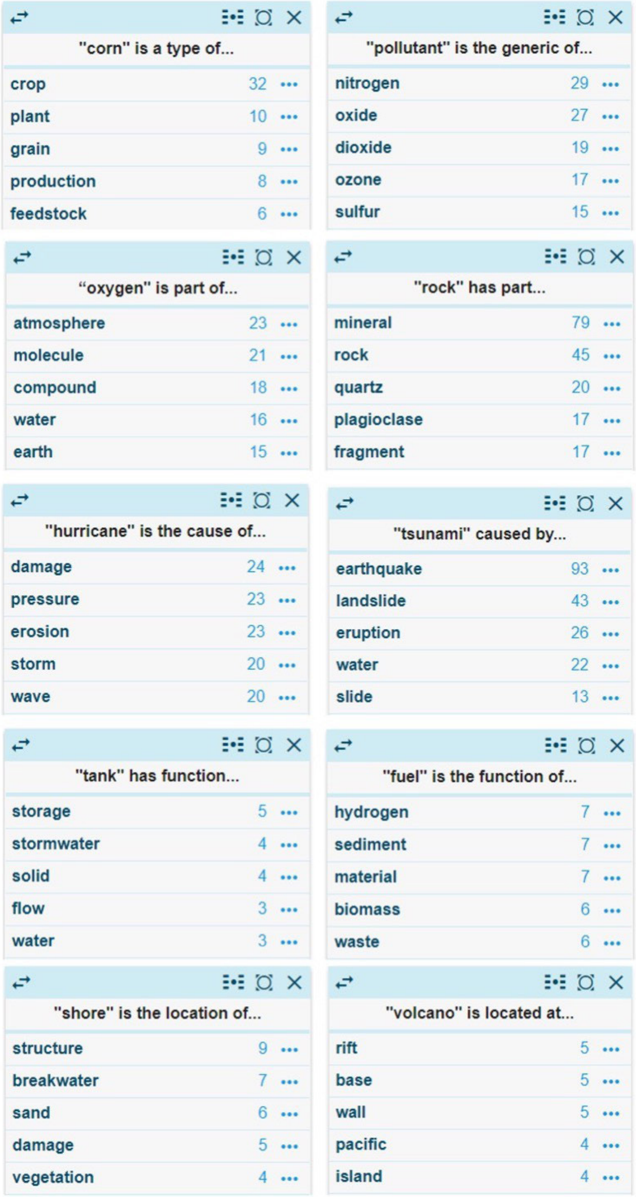
2 Method

This adaptation of the default English sketch grammar currently envisages the following: (i) creation of new gramrel; (ii) splitting and merging of gramrels; (iii) modification of gramrel; (iv) suppression of gramrels; and (v) a combination of these strategies. This paper focuses on the creation and evaluation of a new gramrel called “X is the proto-agent of…/X is the proto-patient of…” as well as the splitting, merging, and modification of the gramrels “modifiers of X” and “adjective predicates of X”.

For this purpose, we applied a modified version of the methodology of creating knowledge- pattern-based sketch grammars (San Martín et al., 2020, p. 5954). This methodology is based primarily on the iterative refinement and evaluation of CQL rules. The corpus (7,249,297 words) that we used consisted of specialized texts on Agronomy from the following sources:

- 36.7 %: theoretical and practical documents on Agronomy published by the Food and Agriculture Organization (FAO) and various national and regional governments in English-speaking countries.
- 30.1 %: specialized monographs and encyclopedias on Agronomy.
- 22.6 %: scientific articles from the International Journal of Agronomy.
- 10.6 %: articles from Wikipedia, manually verified to belong to the field of Agronomy.

In the early stages of gramrel development, the emphasis is on the evaluation of individual rules. Evaluation is performed by querying the rule in a corpus compiled in Sketch Engine and ascertaining whether the rule extracts expected results without generating noise. This type of precision is evaluated in small samples (usually 100 random lines) to verify that the modifications in the rules produce



"corn" is a type of...	
crop	32
plant	10
grain	9
production	8
feedstock	6

"pollutant" is the generic of...	
nitrogen	29
oxide	27
dioxide	19
ozone	17
sulfur	15

"oxygen" is part of...	
atmosphere	23
molecule	21
compound	18
water	16
earth	15

"rock" has part...	
mineral	79
rock	45
quartz	20
plagioclase	17
fragment	17

"hurricane" is the cause of...	
damage	24
pressure	23
erosion	23
storm	20
wave	20

"tsunami" caused by...	
earthquake	93
landslide	43
eruption	26
water	22
slide	13

"tank" has function...	
storage	5
stormwater	4
solid	4
flow	3
water	3

"fuel" is the function of...	
hydrogen	7
sediment	7
material	7
biomass	6
waste	6

"shore" is the location of...	
structure	9
breakwater	7
sand	6
damage	5
vegetation	4

"volcano" is located at...	
rift	5
base	5
wall	5
pacific	4
island	4

Figure 6. Sample of semantic WSs extracted with the ESSG from the EcoLexicon English corpus

the expected result. Since evaluating recall would slow down the process, the number of total matches extracted by the rule is taken as a proxy for recall (i.e., the greater the number of matches, the greater the recall).

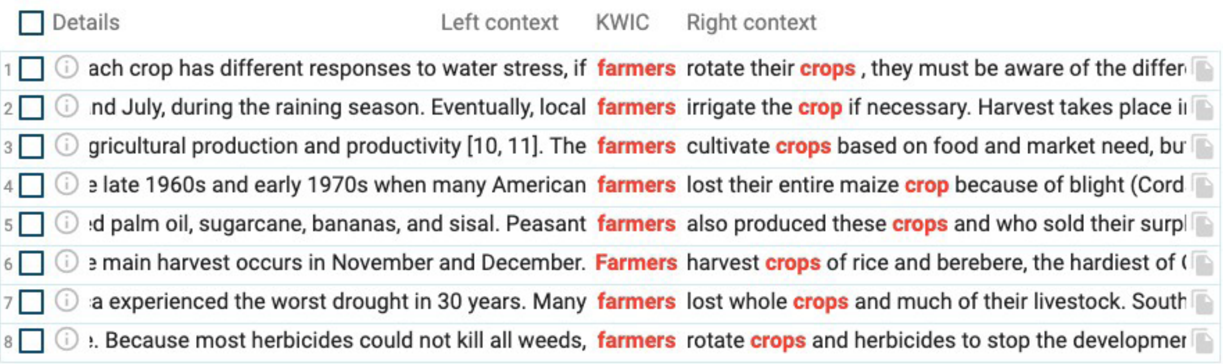
For sketch grammars, the precision of rules is less important than recall. Since users access results in the form of WS (lists of results ordered by frequency or association score), exceptions, errors, and other noisy results tend to be relegated to the bottom of WS lists. More frequent and significant results tend to appear at the top. For this reason, during the development of gramrels, it is also important to periodically test the resulting WS, even though this is more time-consuming because the sketch grammar needs to be previously compiled. Accordingly, this research study evaluated our two adaptations of the default sketch grammar on the basis of the results in WS form (see section 2.3 for the evaluation methodology).

The gramrels can be downloaded at <https://uqtr.ca/knowledge-sketch-grammar/>. Instructions on how to use them in Sketch Engine are also available at that address.

2.1 The proto-agent-patient gramrel

The organization of specialized domains is based on events in which the interaction between different types of agents and patients plays a predominant role. (Faber, 2015, p. 23). However, it is not currently possible to extract the agent-patient relation in Sketch Engine in a user-friendly way. For this reason, we developed a new gramrel that extracts the relation between the nouns functioning as subject and object in the same sentence (e.g., *farmer* and *crop* in Figure 7). This syntactic relation is useful for specialized knowledge extraction because the subject usually accomplishes an action that affects the object in some way. In terms of semantic roles, the former is normally characterized as the agent, but depending on the verb, it can also be an experiencer, an instrument, among other roles. The object can be typically labeled as the patient, but also as the theme, the recipient, among other roles. Dowty (1991) groups these semantic roles into two macroroles: proto-agent and proto-patient. Consequently, this gramrel is called "X is the proto- agent of…/X is the proto-patient of…" (proto-agent-patient gramrel).

The first step consisted in creating a basic version of the gramrel by combining the two default gramrels "objects of "X"/verbs with "X" as object" (object gramrel) and "subjects of "X"/verbs with "X" as subject" (subject gramrel). New rules are thus created by combination instead of



	□ Details	Left context	KWIC	Right context
1	□ ⓘ	ach crop has different responses to water stress, if	farmers	rotate their crops , they must be aware of the differ
2	□ ⓘ	nd July, during the raining season. Eventually, local	farmers	irrigate the crop if necessary. Harvest takes place i
3	□ ⓘ	gricultural production and productivity [10, 11]. The	farmers	cultivate crops based on food and market need, bu
4	□ ⓘ	e late 1960s and early 1970s when many American	farmers	lost their entire maize crop because of blight (Cord
5	□ ⓘ	d palm oil, sugarcane, bananas, and sisal. Peasant	farmers	also produced these crops and who sold their surpl
6	□ ⓘ	e main harvest occurs in November and December.	Farmers	harvest crops of rice and berebere, the hardiest of (
7	□ ⓘ	a experienced the worst drought in 30 years. Many	farmers	lost whole crops and much of their livestock. South
8	□ ⓘ	. Because most herbicides could not kill all weeds,	farmers	rotate crops and herbicides to stop the developmer

Figure 7. Concordances with *farmer* as subject and *crop* as object in our Agronomy corpus

grouping the corresponding rules in a single gramrel. The active-voice rules are combined in a new rule, whereas the passive-voice rules are merged into another rule (Figure 8).

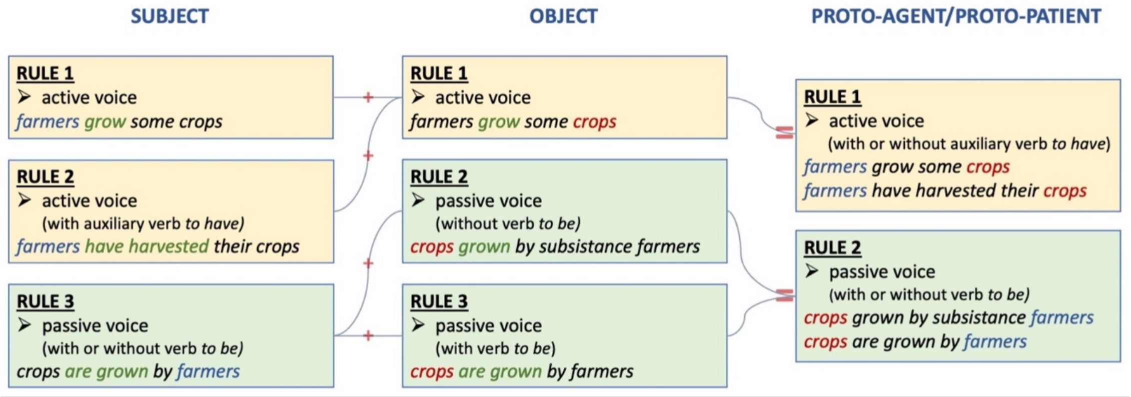

SUBJECT
OBJECT
PROTO-AGENT/PROTO-PATIENT
RULE 1
➢ active voice
farmers grow some crops
RULE 2
➢ active voice
(with auxiliary verb to have)
farmers have harvested their crops
RULE 3
➢ passive voice
(with or without verb to be)
crops are grown by farmers
RULE 1
➢ active voice
farmers grow some crops
RULE 2
➢ passive voice
(without verb to be)
crops grown by subsistance farmers
RULE 3
➢ passive voice
(with verb to be)
crops are grown by farmers
RULE 1
➢ active voice
(with or without auxiliary verb to have)
farmers grow some crops
farmers have harvested their crops
RULE 2
➢ passive voice
(with or without verb to be)
crops grown by subsistance farmers
crops are grown by farmers
+
+
+
+
=
=

Figure 8. Combination of subject and object rules into the proto-agent-patient rules

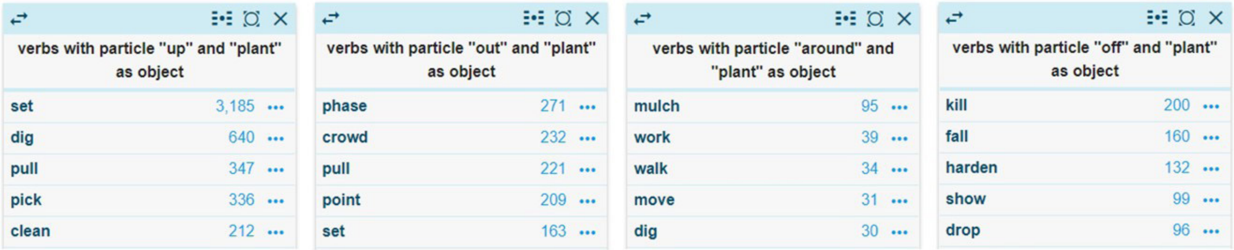
This basic version was used as a benchmark in the evaluation. Since it only returns 80,814 matches in our corpus, it was enriched and refined to increase recall. Some of the changes that allowed us to increase recall without compromising precision included the following:

- Optional modal verbs (*will*, *can*, *must*, etc.): "…any ammonium-containing fertilizer *will* ultimately decrease soil pH…".
- Additional optional auxiliary verbs: "…wind erosion *is* causing significant soil loss…".
- Possibility of certain subordinate structures (is capable of…, have the advantage/ ability/… of/to, seems/appears/… to…, is used/designed/intended to…, etc.): "…parasitic nematodes *are capable of* causing plant diseases...", "…cover crops *are used to* improve the soil structure...".

We also created a new rule that captures the subject-object relation that could not be derived from the subject and object gramrels: PROTO-PATIENT that PROTO-AGENT affects (and variants). This rule matches concordances such as "…nitrogen that the crop roots can take up and use" or "… management practices that farmers adopt focus on herbicides".

Once these changes were applied, the resulting rules were evaluated and three limitations were identified: phrasal verbs, multiple nouns in subject or object position, and verbs that do not convey a proto-agent-patient relation or inverse the order (since the subject is a proto-patient, and the object, a proto-agent).

Regarding phrasal verbs, the basic version does not allow the presence of a preposition between the verb and the object. Instead, Sketch Engine's default sketch grammar displays phrasal verbs in specific columns (Figure 9). Since not including them in the proto-agent-patient gramrel reduces recall, we allowed up to two optional prepositions between the verb and the object to retrieve concordances such as "…when the crop takes up most of the nitrogen…" or microbes break down organic matter". Even though this occasionally generates noise (e.g., "…goods travel from manufacturers to distributors…"), preliminary tests showed that the increase in recall compensated for it.



verbs with particle "up" and "plant" as object	
set	3,185
dig	640
pull	347
pick	336
clean	212

verbs with particle "out" and "plant" as object	
phase	271
crowd	232
pull	221
point	209
set	163

verbs with particle "around" and "plant" as object	
mulch	95
work	39
walk	34
move	31
dig	30

verbs with particle "off" and "plant" as object	
kill	200
fall	160
harden	132
show	99
drop	96

Figure 9. Phrasal verbs WS columns of *plant* in the corpus enTenTen18

A second problem was that WS results can only be single words. The subject and object gramrels only capture the last noun in noun compounds (e.g., "…nitrogen applications *disturb* the soil". In the case of noun compounds linked by a preposition (e.g., "Rotation *of* crops increases the production *of* biomass...") and enumerations (e.g., "Cattle digestion, fertilizers and animal wastes cause emissions..."), only the closest noun to the verb is captured. These limitations are justified in that they protect the precision of the rules. However, they limit recall because the noun compound head is not always detected. (e.g., "The pollution of surface and groundwater that produces serious health problems…"). Likewise, some nouns do not occupy the subject or object head position but semantically could act as proto-agent or proto-patient. For instance, in "…nitrogen applications disturb the soil", the head of the subject is *applications* and is, therefore, the direct proto-agent. However, *nitrogen* is indirectly a proto-agent as well.

Therefore, we modified the rules to capture any noun (whether head or modifier) in a nominal compound or an enumeration. This included both noun compounds without a preposition (e.g., "hydrocarbon pesticide residue") or linked by a preposition *of* (e.g., "fixation of nitrogen"). No other prepositions were included at this point because preliminary evaluations showed that they were an important source of noise. Additionally, we enabled the rule to capture all the nouns in enumerations either in subject or object position (e.g., "These fungi infect many cereals, grasses and other plants..."). We also included enumerations with hyponymic formulas (e.g., "…organisms such as fungi and nematodes can damage…").

The third limitation concerns the fact that the subject-object relation does not always correspond to the proto-agent-patient relation. This was addressed by filtering certain verbs. A first group of verbs (invalidating verbs) are those that do not convey the proto-agent-patient relation, for example, *to be*. This group also includes verbs that convey certain relations already captured with the semantic WS, such as hyponymy (e.g., *include*) or meronymy (e.g., *have*). The second group includes those that invert the common argument order (inverting verbs). Table 1 includes both lists of verbs. While invalidating verbs were excluded from all the rules, the rules were duplicated for inverting verbs. In one set of rules, the inverting verbs were excluded, and in the others, the position of the proto-agent and proto-patient were interchanged.

Table 1. Invalidating and inverting verbs

invalidating verbs	accord, arise, be, become, belong, come, compete, compose, comprise, consist, contain, define, exist, feature, follow, gain, happen, have, include, lack, lose, match, name, need, originate, range, receive, refer, regard, relate, remain, require, stay, stem, survive, vary
inverting verbs	depend (on), rest (on), lie (on), lie (in), result (from), suffer (from), rely (on/upon)

The enriched version of the gramrel returned 374,525 matches from our corpus, four times more than the basic version. It is composed of the following six rules (note that the verb *affect* represents any verb with the above-mentioned exceptions):

- PROTO-AGENT affects PROTO-PATIENT (and variants)
- PROTO-PATIENT is affected by PROTO-AGENT (and variants)
- PROTO-PATIENT that PROTO-AGENT affects (and variants)
- PROTO-AGENT affects PROTO-PATIENT (and variants) (inverting verbs)
- PROTO-PATIENT is affected by PROTO-AGENT (and variants) (inverting verbs)
- PROTO-PATIENT that PROTO-AGENT affects (and variants) (inverting verbs)

2.2 Adjectival gramrels

Two WS columns in the default sketch grammar extract the adjectives that modify a given noun. The "modifiers of X/nouns modified by X" gramrel (modifiers gramrel) retrieves the adjectives and nouns that appear before a noun (e.g., "perennial crop", "corn crop"). The "adjectives predicates of X/subjects of be X" gramrel (predicative adjectives gramrel) extracts an adjective placed after a noun even when separated by *to be* (e.g., "crops resistant to herbicides", "crops are tolerant").

These columns have potentially useful features for building definitions or conceptual networks, namely the ability to assign characteristics to a given concept. To adapt them to the extraction of specialized knowledge, the modifiers gramrel was divided in two: one gramrel for adjectives modifiers and another for noun modifiers. This allowed us to explore whether a single column for all the adjectives that modify a noun was useful. In another gramrel, all the nouns modifying another noun (either preceding the modified noun or postposed with a preposition, e.g., "wheat production" and "production of wheat") could also be grouped. This paper only addresses adjectives from the point of view of a noun search word. In other words, whereas the adjectives gramrel is dual ("adjectives of X/nouns modified by X"), it only focuses on the column that lists adjectives ("adjectives of X").

This new adjectives gramrel is composed of three gramrels that will be tested separately: the attributive adjectives gramrel, predicative adjectives gramrel, and the hyponymic adjectives gramrel.

The attributive adjectives gramrel extracts the adjectives that precede the noun modified (e.g., "synthetic fungicide", "foliar fungicide"). This gramrel originates from the split of the modifiers gramrel. It is composed of a single rule, which was changed to exclude the following adjectives that are not useful for specialized knowledge extraction: *most, least, many, other, more, less, such, able, unable, due, capable, incapable, various, several, few, same, different*. These adjectives were also excluded from the other adjectival gramrels. This gramrel produced 514,578 matches in our corpus.

The predicative adjectives gramrel is composed of a single rule in the default grammar. To create an enriched version, we divided it into two separate rules. The first rule captures the adjective placed directly after the noun (e.g., "…keep the soil dry…") and was modified so as to capture two adjectives (e.g., "…fisheries more productive and sustainable…"). The second rule extracts the adjective placed after the noun and the verb *to be* (e.g., "the soil is dry"). We increased its recall by adding more predicative verbs (i.e., *appear*, *look*, *seem*, *become*, *remain*, *get*, *turn*). Other modifications include optional auxiliary verbs (e.g., "droughts have become more prolonged"), modal verbs (e.g., "soybeans will remain yellow"), two adjectives (e.g., "soil is acid or alkaline"), and noun enumerations: (e.g., "leaves and small stems become more brittle"). The basic version of the gramrel produced 34,127 matches in our corpus. The enriched version obtained 46,358 matches, which is a 35.84% increase.

Finally, the hyponymic adjectives gramrel captures the adjectives that qualify the hyponym of the search word. It is based on hyponymic knowledge patterns. It follows the logic that, in a hyponymic structure, the adjective modifying the hypernym potentially expresses a characteristic of the hyponym. For example, in "...the use of interventional measures such as fungicides...", it can be deduced that *fungicide* is a type of *interventional measure*. Therefore, *interventional* also applies to *fungicide*.

The starting point of this gramrel was the hyponymic rules in the ESSG (see 1.2.2.). We only retained those rules that returned at least 1000 results in our corpus. We then excluded the rules that were excessively noisy, although they might be included in the future if they can be refined to yield satisfactory results. The hyponymic adjectives gramrel produced 27,420 matches in our corpus and contains the following rules:

1. *adjective* HYPERNYM such as/including/especially/like/includes HYPONYM (e.g., "…agricultural inputs such as herbicides")
2. HYPONYM and/or other *adjective* HYPERNYM (e.g., "…antibiotics or other effective antimicrobials)

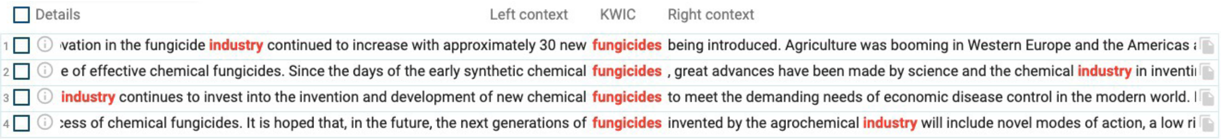
2.3 Evaluation methods

These adaptations were evaluated in two stages. The first stage evaluated the WS columns in terms of precision, whereas the second stage evaluated the usefulness of the gramrels to extract specialized knowledge.

In both stages, the evaluation was performed using one high-frequency term (*crop* with 28,457 occurrences in the Agronomy corpus) and two medium-frequency terms (*fungicide* with 1,161 occurrences, and *nematode* with 866 occurrences) as search words. Additionally, for the second stage, we extracted definitions of these terms from specialized glossaries and multidomain terminology databases (only if the definition was labeled as belonging to Agronomy or its subdomains). In total, 30 definitions of *crop*, 20 of *nematode*, and 26 of *fungicide* were recovered. In all stages, only the first five most frequent results per WS column were considered. To evaluate the precision (i.e., the percentage of correct results), we assessed whether the results were correct by accessing the corresponding concordance lines. In the case of the proto-agent-patient gramrel, a concordance line was considered correct (i.e., a true positive) if a proto-agent-patient relationship can be deduced directly or indirectly from the concordance.

For attributive and predicative adjectives, a concordance line was considered correct if the adjective qualified the captured noun in the concordance. In the case of hyponymic adjectives, it was considered correct if the noun inherited the characteristic expressed by the adjective.

We also calculated validity, according to which a result in a WS column is valid when at least one of its associated concordances is a true positive (San Martín et al., 2020, p. 5961). For example, the relation "*fungicide* is the proto-patient of *industry*" has four associated concordances (Figure 10). Since only 3 and 4 are correct, the precision of this result is 50%. However, because there is at least one correct concordance, validity is 100%.



	□ Details	Left context	KWIC	Right context
1	□ ⓘ	vation in the fungicide industry continued to increase with approximately 30 new	fungicides	being introduced. Agriculture was booming in Western Europe and the Americas ;
2	□ ⓘ	e of effective chemical fungicides. Since the days of the early synthetic chemical	fungicides	, great advances have been made by science and the chemical industry in inventi
3	□ ⓘ	industry continues to invest into the invention and development of new chemical	fungicides	to meet the demanding needs of economic disease control in the modern world. I
4	□ ⓘ	:ess of chemical fungicides. It is hoped that, in the future, the next generations of	fungicides	invented by the agrochemical industry will include novel modes of action, a low ri

Figure 10. Concordances associated to the relation "*fungicide* is the proto-patient of *industry*"

The second evaluation stage explored the usefulness of the adaptations for specialized knowledge extraction by comparing the gramrels results with the definitions of the search terms. In some cases, they were also compared with the contextonymic and semantic WS columns.

3 Results

3.1 Evaluation of the proto-agent-patient gramrel

3.1.1 First stage

The table with the complete results of the precision and validity analysis is in Appendix 1. Figure 11 summarizes the results in terms of precision and validity, and Figure 12 reproduces the resulting WS columns. On average, the enriched version has a precision of 71.17% compared to 66.01% for the basic version. The enriched version was found to perform better for *fungicide* and *crop*, whereas the basic version performed better for *nematode*.

In terms of validity, the results of the enriched version are also slightly higher than the basic version. On average, the enriched one obtained 93.33% validity, while the basic one obtained 83.33%. The enriched version performed better for *nematode* and *fungicide* than the basic version. Both versions were 100% valid for *crop*.

The analysis of the incorrect concordances allowed us to identify different types of errors, mostly attributable to the limitations of WSs. Such errors include problems with sentence segmentation (e.g., "Destroy or control weeds and soil pests Incorporate crop residues...") and POS-tagging (e.g., "This water blistering disorder crops up from time to time..."). This type of error was present in the same proportion in the basic and enriched versions.

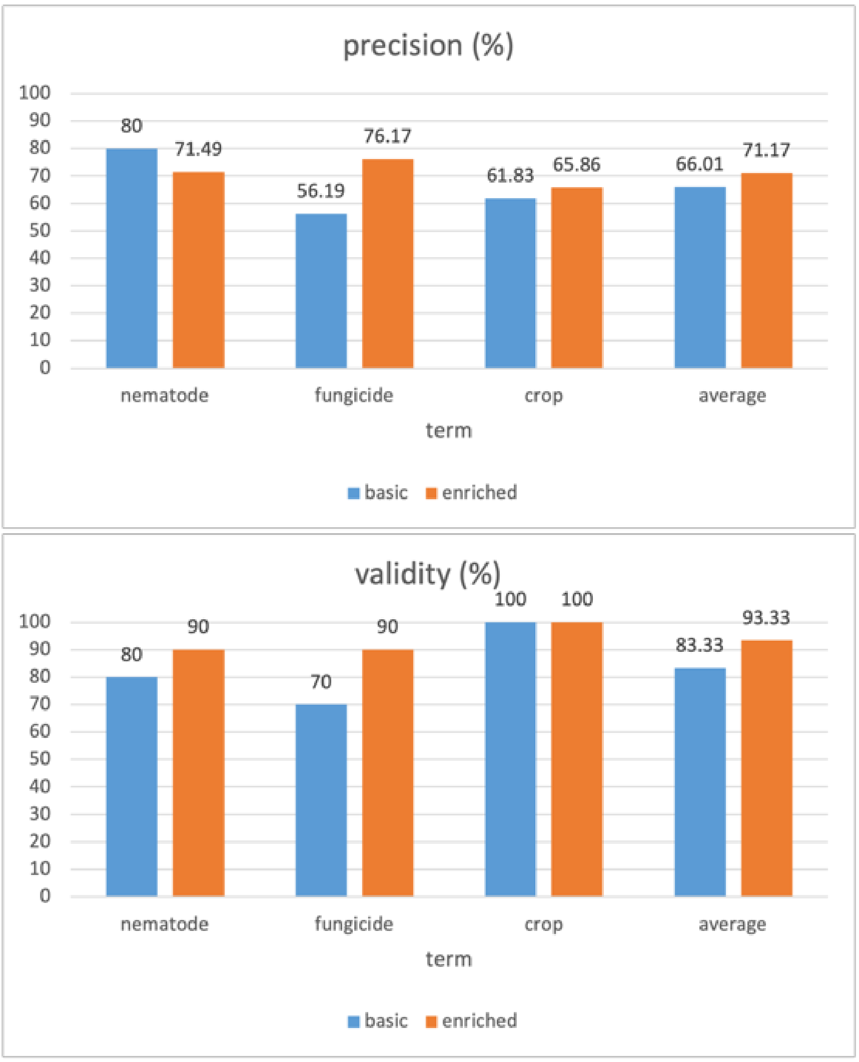
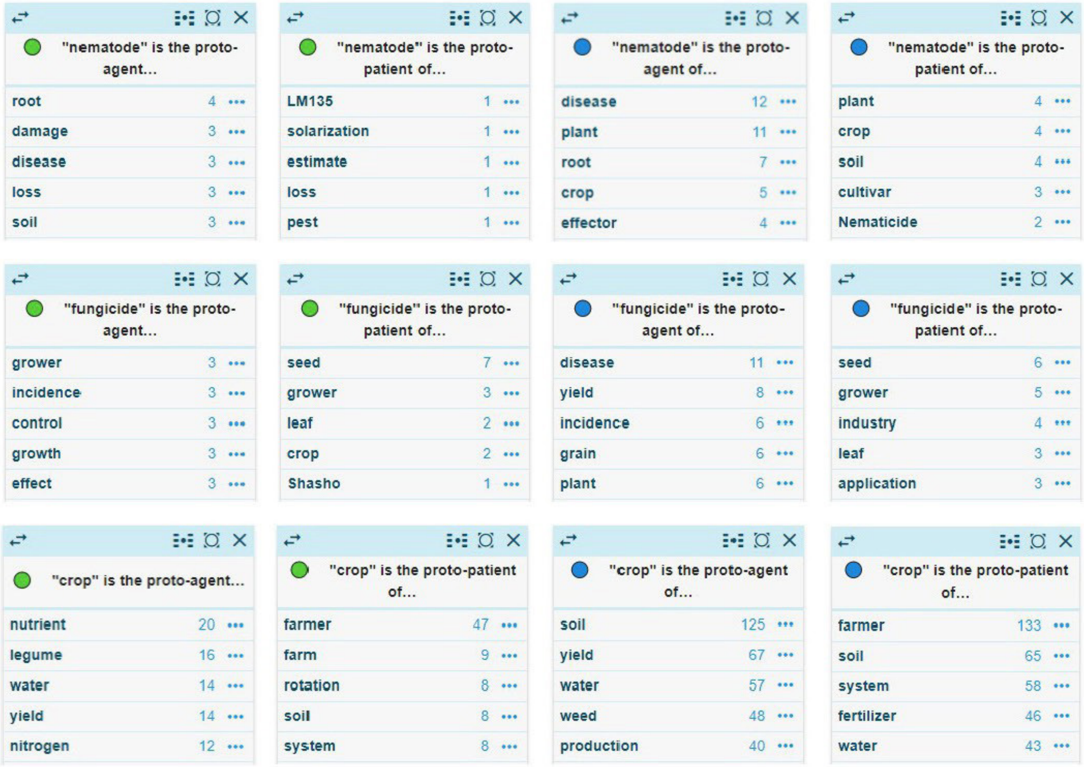


Figure 11. Precision and validity of the proto-agent-patient gramrel

Another type of error caused by a WS limitation concerned noun compounds. As previously explained, to overcome this problem, the enriched version retrieves all the nouns in nominal compounds. Although this facilitates the retrieval of many correct results, it is also a source of noise (e.g., "Plant pathologists are investigating methods of nematode control". However, these preliminary results seem to indicate that the increase in recall compensates for the noise, especially since it is preferable to prioritize recall rather than precision.

Finally, there were errors generated by invalidating and inverting verbs. Even though the enriched version accounts for some of these verbs, new ones appeared in the concordances (e.g., "Crops may *tolerate* greater amounts of blowing soil..."). All these verbs will be analyzed before including the rules.



● "nematode" is the proto-agent...	
root	4
damage	3
disease	3
loss	3
soil	3

● "nematode" is the proto-patient of...	
LM135	1
solarization	1
estimate	1
loss	1
pest	1

● "nematode" is the proto-agent of...	
disease	12
plant	11
root	7
crop	5
effector	4

● "nematode" is the proto-patient of...	
plant	4
crop	4
soil	4
cultivar	3
Nematicide	2

● "fungicide" is the proto-agent...	
grower	3
incidence	3
control	3
growth	3
effect	3

● "fungicide" is the proto-patient of...	
seed	7
grower	3
leaf	2
crop	2
Shasho	1

● "fungicide" is the proto-agent of...	
disease	11
yield	8
incidence	6
grain	6
plant	6

● "fungicide" is the proto-patient of...	
seed	6
grower	5
industry	4
leaf	3
application	3

● "crop" is the proto-agent...	
nutrient	20
legume	16
water	14
yield	14
nitrogen	12

● "crop" is the proto-patient of...	
farmer	47
farm	9
rotation	8
soil	8
system	8

● "crop" is the proto-agent of...	
soil	125
yield	67
water	57
weed	48
production	40

● "crop" is the proto-patient of...	
farmer	133
soil	65
system	58
fertilizer	46
water	43

Figure 12. Resulting WS columns from the proto-agent-patient gramrel. The basic version has a green dot. The enriched version, a blue dot

3.1.2 *Second stage*

Since the enriched version returned more precise results, the second stage was performed on this one. First, we compared the proto-agents and proto-patients in the extracted definitions of the analysis terms (Table 2) with the gramrel results (Table 3). Terms present in both the definitions and the WS are in bold.

Table 2. Proto-agents and proto-patients in the definitions. The number of occurrences in the definitions is in parentheses

fungicide	is the proto-agent of…	fungus(24), growth(6), **disease(3)**, **plant(3)**, mold(2), mildew(2), control(1), yeast(1), pathogen(1), crop(1), product (1), soil(1), development(1)
	is the proto-patient of…	insect(1), ant(1)
crop	is the proto-agent of...	-
	is the proto-patient of...	livestock(1), labor(1), **farmer(1)**, people(1)
nematode	is the proto-agent of...	**plant(11)**, **root(5)**, animal(5), **crop(2)**, vine(2), tissue(2), agriculture(1), pest(1), slug(1), leatherjacket(1), loss(1), human(1), yield(1), swelling(1), growth(1), **disease(1)**, bird(1), mammal(1), insect(1), juice(1), structure(1), damage(1)
	is the proto-patient of...	plant(1), control(1), brassica(1), chemical(1), **nematicide(1)**, water(1), contamination(1)

Table 3. Top five results in the proto-agent-patient WS with the number of correct concordances in parentheses

fungicide	is the proto-agent of…	**disease(10)**, yield(7), incidence(6), grain(6), **plant(4)**
	is the proto-patient of…	grower(5), industry(2), leaf(3), application(2), ~~seed(0)~~
crop	is the proto-agent of...	soil(92), water(38), yield(37), weed(35), production(26)
	is the proto-patient of...	**farmer(122)**, soil(40), system(37), water(25), fertilizer(23)
nematode	is the proto-agent of...	**disease(10)**, **plant(10)**, **root(6)**, **crop(4)**, effector(4)
	is the proto-patient of...	crop(3), cultivar(3), soil(2), **nematicide(1)**, ~~plant(0)~~

Of the 28 valid WS results, only 8 (i.e., 28.57%) appear as well in the definitions. This low percentage was to be expected because definitions select the conceptual information considered most relevant. Corpora tend to contain much more information. It can also be observed that the use of the proto-agent-patient relation in the definitions is variable: *fungicide* and *nematode* are most frequently defined as proto-agent, while in the definitions of *crop*, both macroroles are rare.

Some frequent proto-agents in the definitions are missing in the WS results. The most prominent case is *fungus* as a proto-agent of *fungicide*. In the column "*fungicide* is the proto-agent of...", *fungus* is in 85th position with only one associated concordance ("...fungicides kill fungi..."). The contextonymic WS of *fungicide* was thus consulted to determine whether the low result of *fungi* was a case of silence (i.e., that the proto-agent-patient gramrel missed relevant concordances).

Fungus is only the 72nd contextonym of *fungicide* with 96 concordances. This indicates that, although the relationship between *fungicide* and *fungus* is relevant to define *fungicide*, specialized texts do not often mention this characteristic. Additionally, we observed that only 12 out of the 96 concordances directly or indirectly convey a proto-agent-proto-patient relationship. However, in most cases, it is not reflected in a subject-object relation (e.g., "…strobilurin fungicides are very active against many plant pathogenic fungi…"). Among the ones in which there is indeed a subject- object relation, most are cases of anaphora (e.g., "…fungicides penetrate into plant tissue, where they kill or inhibit a fungus…") and uncommon use of punctuation (e.g., "…fungicides (kill fungi)..." Both cases are difficult to account for in CQL rules without generating excessive noise.

Finally, regarding the complementarity of this relationship with the contextonymic WS, the first contextonymic results of the three terms show matches with the proto-agent-patient gramrel. It follows that the proto-agent-patient columns can facilitate the discovery of the relationship between the search word and many of its contextonyms. As shown in Table 4, for 9 of the 15 most frequent contextonyms of *fungicide*, *nematode*, and *crop*, the first five proto-agent-patient results help to determine the semantic relationship between the two terms.

Table 4. Comparison of contextonyms (number of occurrences in parenthesis) and the results from the proto-agent-patient gramrel

search word	**contextonym**	**fungicide is its…**
fungicide	disease(907)	proto-agent
	crop(581)	proto-agent/proto-patient
	application(462)	-
	resistance(421)	-
	control(406)	-
nematode	plant(776)	proto-agent
	soil(600)	proto-patient
	crop(398)	proto-agent
	root(347)	proto-agent
	population(261)	-
crop	soil(15,672)	proto-agent
	plant(9,063)	-
	production(8,707)	proto-agent
	use(verb)(7,944)	-
	yield(7,465)	proto-agent

3.2 Evaluation of the adjectives gramrel

3.2.1 First stage

The table with the complete results of the precision and validity analysis is in Appendix 2. Figure 13 summarizes the precision and validity data. Figure 14 reproduces the WS columns.

As for the attributive adjectives gramrel, it has a very high precision (99.93 % of average) and a perfect validity. The only errors that the gramrel produced are due to problems with corpus segmentation.

As for predicative adjectives, the enriched version performed slightly better than the basic version in both precision (72.53% vs. 68.47 %) and validity (80% vs. 75.94%). It is not surprising that *crop*, the most frequent term, obtained the best results. In fact, in the enriched version, the validity is 100%. Despite the small sample, the precision and validity are sufficiently high to conclude that this gramrel performs satisfactorily. Moreover, the results indicate that the enriched version, even retrieving 35.84% more results, maintains or even surpasses the level of precision of the basic version.

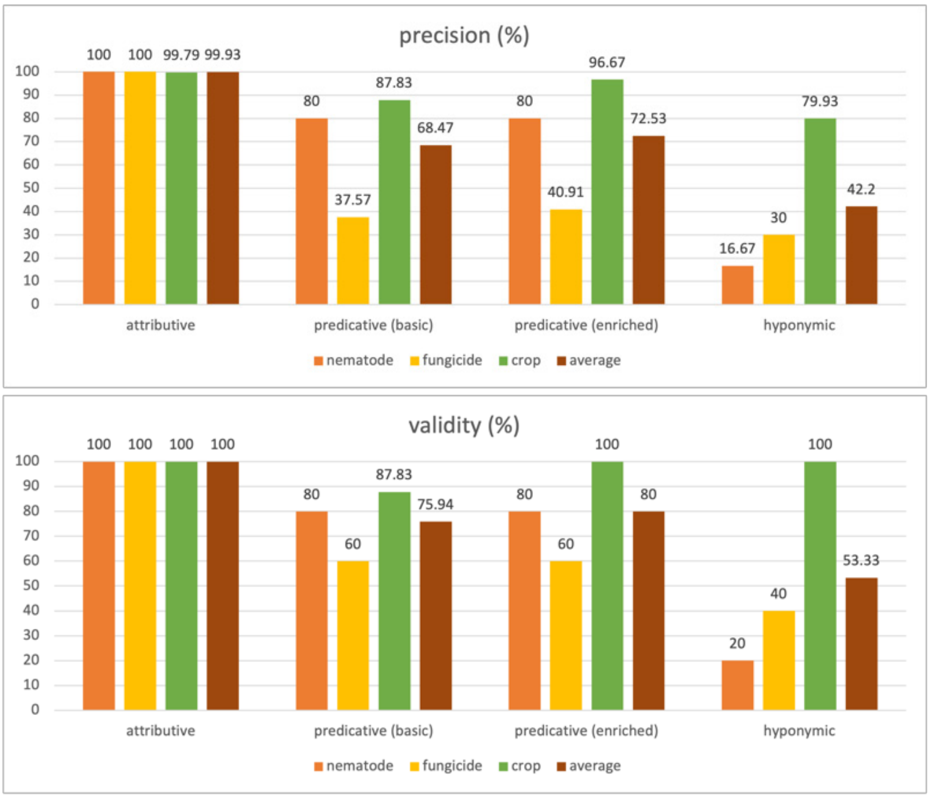


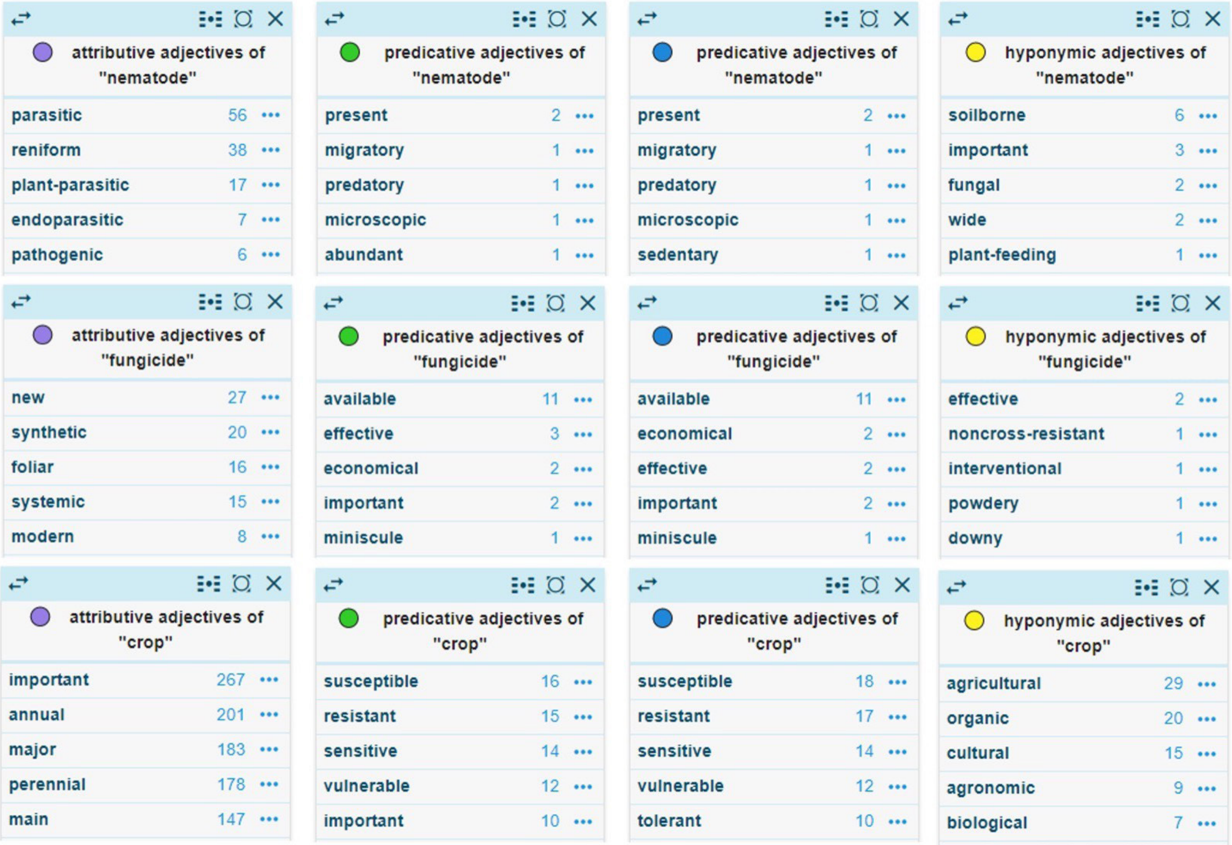
Figure 13. Precision and validity of the of the adjectival gramrels

As for the frequent errors observed in the extraction rules of predicative adjectives, they are similar to the ones of the proto-agent-patient gramrel. Both the basic and enriched versions retrieve some incorrect results stemming from problems with POS tagging and corpus segmentation. The enriched version also matches some incorrect matches due to noun compounds: "...combinations of these fungicides are effective in the management…".

Finally, the results of the hyponymic gramrel are notably inferior to the other gramrels both in precision and validity. While the precision is 42.2%, the validity is only slightly higher at 53.33%. However, the results for *crop* are 79.93% precise and 100% valid. Since crop is the most frequent term, this is consistent with the fact that knowledge-pattern-based gramrels (such as this gramrel) need larger corpora to yield satisfactory results.

This gramrel shares the same errors as the other ones, but there are also two that are unique to it. The first error occurs when the head of the hypernym is a collective noun. For example, in "there is a wide variety of parasites including trematodes, cestodes, nematodes…", *wide* qualifies *variety*, and not *nematode*'s hypernym (*parasite*). Therefore, *nematode* cannot be said to inherit that attribute. This error can easily be solved by filtering out collective nouns.

The other error occurs when the head of the hypernym and the head of the hyponym are the same lemma. For example, "…plant-feeding nematodes such as lesion nematodes". In this case, *plant- feeding* only applies to one particular type of nematode (i.e., *lesion nematodes*). This can also be easily filtered to improve the performance of this gramrel. However, the attributive adjectives gramrel already captures the adjective preceding the hypernym (i.e., "plant-feeding nematodes".)



● attributive adjectives of "nematode"	
parasitic	56
reniform	38
plant-parasitic	17
endoparasitic	7
pathogenic	6

● predicative adjectives of "nematode"	
present	2
migratory	1
predatory	1
microscopic	1
abundant	1

● predicative adjectives of "nematode"	
present	2
migratory	1
predatory	1
microscopic	1
sedentary	1

○ hyponymic adjectives of "nematode"	
soilborne	6
important	3
fungal	2
wide	2
plant-feeding	1

● attributive adjectives of "fungicide"	
new	27
synthetic	20
foliar	16
systemic	15
modern	8

● predicative adjectives of "fungicide"	
available	11
effective	3
economical	2
important	2
miniscule	1

● predicative adjectives of "fungicide"	
available	11
economical	2
effective	2
important	2
miniscule	1

○ hyponymic adjectives of "fungicide"	
effective	2
noncross-resistant	1
interventional	1
powdery	1
downy	1

● attributive adjectives of "crop"	
important	267
annual	201
major	183
perennial	178
main	147

● predicative adjectives of "crop"	
susceptible	16
resistant	15
sensitive	14
vulnerable	12
important	10

● predicative adjectives of "crop"	
susceptible	18
resistant	17
sensitive	14
vulnerable	12
tolerant	10

○ hyponymic adjectives of "crop"	
agricultural	29
organic	20
cultural	15
agronomic	9
biological	7

Figure 14. Resulting WS columns from the adjectival gramrels. The basic version of the predicative adjectives has a green dot. The enriched version, a blue dot.

3.2.2 Second stage

In this stage, we left out the basic version of the predicative adjectives gramrel and evaluated the attributive and hyponymic adjectives gramrel along with the enriched predicative adjectives gramrel. First, we compared the gramrel results (Table 5) with the adjectives extracted from the definitions of the terms under analysis (Table 6). In both tables, the terms present in both the definitions and WSs are in bold.

Table 5. Adjectives in the definitions. The number of occurrences in the definitions is in parentheses

fungicide	chemical (6), physical (2), toxic (1)
nematode	**microscopic (7)**, **parasitic (6)**, cylindrical (3), small (3), unsegmented (3), slender (2), elongated (2), worm-like (2), nonsegmented (2), numerous (1), multicellular (1), living (1), beneficial (1), harmful (1), former (1), biological (1), phytoparasitic (1), abundant (1), long (1), legless (1), worm-shaped (1), aquatic (1), round (1), colorless (1), threadlike (1)
crop	cultivated (9), grown (8), **agricultural (2)**, used (2), growing (2), young (2), managed (2), horticultural (1), animal (1), total (1), yearly (1), wild (1)

Table 6. Top five results in the adjectival WSs with the number of correct concordances in parentheses

	attributive	**predicative**	**hyponymic**
fungicide	new (27) synthetic (20) foliar (16) systemic (15) modern (8)	available (6) important (2) effective (1) ~~economical (0)~~ ~~miniscule (0)~~	effective (1) interventional (1) ~~powdery (0)~~ ~~noncross-resistant (0)~~ ~~downy (0)~~
nematode	**parasitic (56)** reniform 38 plant-parasitic (17) endoparasitic (7) pathogenic (6)	present (2) migratory (1) **microscopic (1)** sedentary (1) ~~predatory (0)~~	soilborne (5) ~~important (0)~~ ~~wide (0)~~ ~~fungal (0)~~ ~~plant-feeding (0)~~
crop	important (267) annual (200) major 183) perennial (178) main (147)	resistant (17) susceptible (15) sensitive (14) vulnerable (12) important (10)	**agricultural (28)** cultural (15) organic (13) agronomic (6) biological (5)

The adjectives extracted from the definitions are only the ones that describe the term. For instance, in the definition of *fungicide* as "chemical compound used to control fungi", the adjective *chemical* describes *fungicide* and is therefore included in the list.

These adjectives have a low level of correspondence (3 out of 36, i.e., 8.33%) with the ones extracted with the gramrels. There is also no specific gramrel that has more matches since each one has only one match. This may be due to the fact that corpora tend to contain more information than definitions. In addition, the adjectives that qualify a noun do not necessarily express a defining characteristic of that concept. It may be a characteristic of a subtype or of an instance of the concept. For example, in "…between two periods, the crop is vulnerable to weeds…", the predicative adjective refers to an instance of *crop* and is also time-restricted. Therefore, it cannot be deduced that vulnerability to weeds is a characteristic of crops, only a possible feature. Another example is "Modern fungicides can be applied to crops…". In this case, the attributive adjective *modern* selects a subset of all existing fungicides. Being modern is also a possible characteristic of fungicides though not a defining one.

As for the complementarity of adjectival gramrels and contextonyms, it should be noted that the initial positions of the WS contextonym are usually occupied by nouns. In the case of *crop*, it is necessary to go to the 19th result to find the first adjective (*high*). With *nematode*, the first adjective is *parasitic*, in position 24th. As for *fungicide*, the first adjective (*new*) is in position 18th. As for the semantic gramrels, none of them extract adjectives. For these reasons, an adjectival gramrel does potentially provide complementary information.

4 Analysis and Discussion

The results of the evaluation of the proto-agent-patient gramrel are promising. Despite the small sample, the enriched version was found to achieve over 70% precision and over 90% validity. The gramrel in its current state is thus functional. The evaluation also shows ways to improve the rules. For example, increasing the list of invalidating and inverting verbs will lead to greater precision. A more thorough evaluation of the impact of retrieving all nouns that make up nominal compounds in object and subject position will also allow us to adjust the rules accordingly.

Comparison with the definitions did not yield a high level of correspondence. It was especially relevant that some frequent proto-agents and proto-patients in the definitions were absent from the gramrel results. This was the case of *fungus* in relation to *fungicide*. The analysis of the concordances in which *fungus* is a contextonym of *fungicide* reveals that the addition of patterns other than subject-object to the gramrel could improve the gramrel.

Regarding adjectival gramrels, preliminary results did not reveal whether it would be useful to merge them into a single gramrel. One argument against merging is that the attributive gramrel returned almost seven times as many results as the other two combined, which means that the others would have very little weight in a merged gramrel. However, as yet, there is not sufficient evidence to affirm that they are more useful separately than merged, especially since having too many WS columns could lead to information overload.

As for the usefulness of adjectival WS for specialized knowledge extraction, this preliminary evaluation was inconclusive. None of the gramrels have a higher correspondence with the adjectives extracted from the definitions. However, there is also the question of whether definitions are a good benchmark for evaluating the capacity of WSs to assist in specialized knowledge extraction. New forms of evaluation should thus also be explored.

While there were some very useful adjectives extracted by the gramrels, others were less so because the adjectives that qualify a noun do not always express a defining characteristic. Thus, our results suggest that the adaptation of this gramrel is worth further study. Thanks to the high precision of the attributive and predicative adjectives gramrels, they are a good research tool for that purpose. As for hyponymic adjectives, the results indicate that the rules need to be refined to achieve greater precision before the gramrel can be considered functional.

It is also worth pointing out that we have not evaluated the inverse columns (i.e., using an adjective as a search word to obtain the list of nouns it qualifies). It is probable that these WS columns are useful for studying adjectives from a conceptual point of view.

In future work, we will continue to develop these gramrels and evaluate them with a larger sample. For this purpose, corpora from different specialized domains and different types of search terms will be used. Finally, in parallel, we will further adapt the English default sketch grammar to create a sketch grammar for specialized knowledge extraction that would also include contextonymic and semantic gramrels.

5 References

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, *67*(3), 547. Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 2, pp. 1212–1248). De Gruyter.

Faber, P. (2015). Frames as a Framework for Terminology. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 14–33). John Benjamins.

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (pp. 741–747). Waseda University.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography*, *1*(1), 7–36.

Kilgarriff, A., & Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. *Proceedings of ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, 32–38.

León-Araúz, P., & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"* (pp. 94–99). Globalex.

León-Araúz, P., San Martín, A., & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73–82).

León-Araúz, P., San Martín, A., & Reimerink, A. (2018). The EcoLexicon English Corpus as an open corpus in Sketch Engine. *Proceedings of the 18th EURALEX International Congress* (pp. 893 901). Euralex.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. Bourigault, M.-C. L'homme & C. Jacquemin, (Eds.), *Recent Advances in Computational Terminology* (pp. 279–302). John Benjamins.

San Martín, A. (2016). *La representación de la variación contextual mediante definiciones terminológicas flexibles*. [Doctoral dissertation, University of Granada].

San Martín, A., Trekker, C., & León-Araúz, P. (2020). Extraction of Hyponymic Relations in French with Knowledge-Pattern-Based Word Sketches. *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 5955–5963). ELRA.

San Martín, A. In press. A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*.

Acknowledgements

We thank Natasha Herger, research assistant, for her contribution to the data analysis. This research was carried out as part of project *Développement d'une méthodologie d'élaboration de définitions terminologiques : analyse de corpus et variation contextuelle* (2020-NP-267503) funded by Quebec's Society and Culture Research Fund (FQRSC).

Appendix 1. Results of the proto-agent-patient gramrel evaluation

BASIC VERSION								
Proto-agent	**Proto-patient**	**Correct/ total**	**Prec. (%)**	**Most common verb**	**Prec. (%)**	**Val. (%)**	**Prec. (%)**	**Val. (%)**
nematode	root	4/4	100	invade	100	100	**80**	**80**
	soil	3/3	100	infest				
	loss	3/3	100	cause				
	disease	3/3	100	cause				
	damage	3/3	100	cause				
LM135	**nematode**	1/1	100		60	60		
solarization		1/1	100					
estimate		1/1	100					
loss		0/1	0					
pest		0/1	0					
fungicide	grower	0/3	0		66.67	80	**56.19**	**70**
	incidence	3/3	100	reduce				
	control	3/3	100	provide				
	growth	3/3	100					
	effect	1/3	33.33					
seed	**fungicide**	2/7	28.57	receive	45.71	60		
grower		3/3	100	use				
leaf		2/2	100	absorb				
crop		0/2	0					
shasho		0/1	0					
crop	nutrient	6/20	30	remove	44.15	100	**61.84**	**100**
	legume	1/16	6.25					
	water	11/14	78.57	use				
	yield	9/14	64.29	produce				
	nitrogen	5/12	41.67	fix				
farmer	**crop**	40/47	85.11	grow	79.52	100		
farm		9/9	100	grow|produce				
rotation		7/8	87.5	involve				
soil		6/8	75	affect				
system		4/8	50	use				

ENRICHED VERSION								
Proto-agent	**Proto-patient**	**Correct/ total**	**Prec. (%)**	**Most common verb**	**Prec. (%)**	**Val. (%)**	**Prec. (%)**	**Val. (%)**
nematode	disease	10/12	83.33	cause	87.99	100	**71.50**	**90**
	plant	10/11	90.91	cause				
	root	6/7	85.71	damage|invade				
	crop	4/5	80	cause				
	effector	4/4	100	secrete				
crop	**nematode**	3/4	75	suppress	55	80		
plant		0/4	0					
soil		2/4	50	affect				
cultivar		3/3	100	support				
Nematicide		1/2	50					
fungicide	disease	10/11	90.91	control	89.02	100	**76.17**	**90**
	yield	7/8	87.5	increase				
	incidence	6/6	100	reduce				
	grain	6/6	100	treat|affect				
	plant	4/6	66.67					
seed	**fungicide**	0/6	0		63.33	80		
grower		5/5	100	use|inquire				
industry		2/4	50					
leaf		3/3	100	absorb				
application		2/3	66.67					
crop	soil	92/125	73.60	improve	66.68	100	**65.86**	**100**
	yield	37/67	55.22	produce|impact				
	water	38/57	66.67	use				
	weed	35/48	72.92	suppress				
	production	26/40	65	improve|increase				
farmer	**crop**	122/133	91.73	grow	65.04	100		
soil		40/65	61.54	affect				
system		37/58	63.79	use				
fertilizer		23/46	50	affect				
water		25/43	58.14	affect				

Appendix 2. Results of the adjectives gramrel evaluation

Gramrel	Term	Adjective	Correct/total	Prec. (%)	Prec. (%)	Val. (%)
attributive adjectives	nematode	parasitic	56/56	100	100	100
		reniform	38/38	100		
		plant-parasitic	17/17	100		
		endoparasitic	7/7	100		
		pathogenic	6/6	100		
predicative adjectives (basic version)		present	2/2	100	80	80
		migratory	1/1	100		
		predatory	0/1	0		
		microscopic	1/1	100		
		abundant	1/1	100		
predicative adjectives (enriched version)		present	2/2	100	80	80
		migratory	1/1	100		
		predatory	0/1	0		
		microscopic	1/1	100		
		sedentary	1/1	100		
hyponymic adjectives		soilborne	5/6	83.33	16.67	20
		important	0/3	0		
		wide	0/2	0		
		fungal	0/2	0		
		plant-feeding	0/1	0		
attributive adjectives	fungicide	new	27/27	100	100	100
		synthetic	20/20	100		
		foliar	16/16	100		
		systemic	15/15	100		
		modern	8/8	100		
predicative adjectives (basic version)		available	6/11	54.55	37.58	60
		effective	1/3	33.33		
		important	2/2	100		
		economical	0/2	0		
		miniscule	0/1	0		
predicative adjectives (enriched version)		available	6/11	54.55	40.91	60
		important	2/2	100		
		effective	1/2	50		
		economical	0/2	0		
		miniscule	0/1	0		
hyponymic adjectives		effective	1/2	50	30	40
		powdery	0/1	0		
		noncross-resistant	0/1	0		
		interventional	1/1	100		
		downy	0/1	0		

attributive adjectives	**crop**	important	267/267	100	99.79	100
		annual	200/201	99.5		
		major	183/183	100		
		perennial	178/179	99.44		
		main	147/147	100		
predicative adjectives (basic version)		susceptible	14/16	87.5	87.83	100
		resistant	15/15	100		
		sensitive	14/14	100		
		vulnerable	11/12	91.67		
		important	6/10	60		
predicative adjectives (enriched version)		susceptible	15/18	83.33	96.67	100
		resistant	17/17	100		
		sensitive	14/14	100		
		vulnerable	12/12	100		
		important	10/10	100		
hyponymic adjectives		agricultural	28/29	96.55	79.93	100
		organic	13/20	65		
		cultural	15/15	100		
		agronomic	6/9	66.67		
		biological	5/7	71.43		

SUNDANESE-MALAY LEXICOGRAPHIC WORKS MANUSCRIPT

Asep Rahmat Hidayat
Regional Agency for Language in West Java Province, Indonesia
asep.rahmat@kemdikbud.go.id

Abstract

Spraeck ende Wordboeck inde Malaysche ende Madagaskarsche Talen met vele Arabische ende Turksche Woorden (1603) by Frederick de Houtman and Vocabularium ofte Woortboeck near order vanden Alphabet in't Duytsch Maleysch ende Maleysche-Duytsch (1623) by Casper Wiltens are two early works of lexicography from the 17th century in Indonesian lexic (1995: xxi, 2000: 7). Just one early Malay lexicographical work had been known before, Raja Ali Haji's Kitab Pengetahuan Bahasa, written in 1858 and published in Singapore in 1927. In addition, Utsman bin Abdullan wrote Kamoes Ketjil Bahasa Melayu-Arab in 1885. Three Sundanese manuscripts are included in the Staatsbibliothek zu Berlin's catalog of manuscripts compiled by Loir and Fathurahman (1999). Meanwhile, according to Pudjiastuti (2018), there were five Sundanese manuscripts in the same library. Two manuscripts are classified as Sundanese-Malay dictionaries on the Staatsbibliothek Zu Berlin's search list (worterbuch: Sundanesisch-Malaiisch). Schoemann IV 2 or PPN 826269338 or Schoemann IV 3 or PPN 826269370 are the two options. Both texts are written in either a Javanese or a Cacarakan script. Schoemann IV 2 was completed on March 7, 1851, according to his colophon. This manuscript predates Kitab Pengetahuan Bahasa by many years. The transcription results and lexicographic analysis of the Schoemann IV 2 manuscript are presented in this paper. The philological method used to convert the cacarakan from Schoemann IV 2 to Latin script. After the manuscript has been transcribed, it will be examined from a lexicographic standpoint.

Keywords: *manuscript**,** lexicographic works**,** Sundanese-Malay*

Introduction

In Indonesia, lexicography history research is still uncommon. Linehan (1949), who recorded lexicographical works from the 16th to the 18th centuries, wrote a brief history of Melayu dictionaries. Only works written by Europeans are included in this study; thus, works written by Malays are not included. The study was then used as a dictionary background guide in the Kamus Besar Bahasa Indonesia (1995) and Sejarah Ejaan Bahasa Indonesia (2000). Proudfoot also lists a number of Malay dictionaries that have been published in Singapore and Malaysia (1993). Meanwhile, lexicographical manuscripts were discovered in many manuscript catalogs. Three Sundanese manuscripts were told at the Staatsbibliothek zu Berlin by Loir and Fatuhurahman (1999). Pudjiastuti (2018) reported that five Sundanese manuscripts were found in the same library. Two manuscripts, Schoemann IV 2 or PPN 826269338 and Schoemann IV 3 or PPN 826269370, are marked as Sundanese-Malay dictionaries on the Staatsbibliothek Zu Berlin website (worterbuch: Sundanesisch-Malaiisch). The manuscript of Schoemann IV 2 was chosen as the subject of this study because it is the first of two similar manuscripts.

Method

The Schoemann IV 2 manuscript is written in *cacarakan* script. *Cacarakan* script is a Javanese or *hanacaraka* script that uses to write the Sundanese language. *Cacarakan* script has eighteen consonants, i.e., *ha, na, ca, ra, ka, da, ta, sa, wa, la, pa, ja, ya, nya, ma, ga, ba, nga*, and seven vowel markers, *i.e., a,* *i, u, e, o, e,* and *eu* (1985).

The Schoemann IV 2 text is transliterated from *cacarakan* script to Latin script. In addition, after the text has existed in Latin script, the vocabulary on the text was then collected according to the theme and presentation in the text.

Results

The *Schoemann IV 2* manuscript contains the number of Sundanese lexicons with the translation in the Malay language. The manuscript contains 137 lexicons of animals, 24 lexicons of household utensils, 11 lexicons of cooking activities, 49 lexicons of types or soil conditions, 50 lexicons of types or house conditions, 45 lexicons of verbs and adjectives, and 12 lexicons of designations or kinship.

Analysis and Discussion

Manuscript

The Schoemann IV manuscript is a part of the *Staatsbibliothek zu Berlin's* collection. There is no detail information on when this manuscript was added to the library's collection. Based on the manuscript code, i.e., Schoemann, this manuscript was likely part of Karl Schoemann's collection. Karl Schoemann (1806-1877) was a German who had lived in Buitenzorg (Bogor) and Batavia (Jakarta) in 1845-1851. He is the children's teacher of the Governor-General of the Dutch East Indies, J.J. Rochussen. He had a passion for culture and was able to collect 350 manuscripts from different parts of Indonesia (2018).

The manuscript is 231 pages long and is printed on European folio paper with the Pro Patria watermark. Thick cardboard is used for the volume, which is lined with brown paper. The paper is 20.5x17 cm in size, with a 12.5x10 cm text area and 9 lines of writing per page. The paper is a light yellowish-white color. The ink used is black, red, and yellow in color. In the text and lighting, black and red are used, while yellow is used in the illumination. The title is written in red ink, the illumination is in the form of flowers and lines, and the writing is in Sundanese. In the Malay language, black ink is used to illuminate buildings with pillars and fences, as well as in Malay literature. Only the first four pages of the book are illuminated.

The text in the manuscript is 107 pages long. The text is missing a few pages. The first 50 pages have been completed. The text jumps to page 125 after page 50. The book is finished from page 125 to page 158. The text jumps to page 194 after page 158. Pages 194 to 212 have been completed. The text jumps to page 226 after page 212. The entire page 230, from page 226 to the end, is complete.

This manuscript's author does not seem to be a Sundanese native speaker. The author is also a well-educated person who is fluent in the Cacarakan script. His confession in the manuscript reveals this: *" beunang kuring diajar hese naker, kuring nulis ieu*"*"* or "boleh saya belajar susah sekali, saya tulis ini" (page 2). This manuscript appears to have been written on request or order by Schoemann himself or others who were Schoemann's informants. " *kuring nyanggakeun kitab ka kangjeng raden, lumayan bae sugan beunang diangge* " or "saya kasih buku kepada tuwan lumayan saja barangkali boleh dipake" is a comment in the manuscript. Even though this manuscript was intended for Schoemann's informant, he was greeted with a polite greeting: *kangjeng Raden or Tuan.*

Bogor is where the manuscript is being written. The words "Nagri Bogor" (page 2) and "nagara Bogor" (page 3) appear in the document (pages 3, 7, 11, 25). The manuscript was completed on March 7, 1851, according to the details in the colophon: "It was already written on Tuesday, 7 Jumadilawal 1266 Hijriah or March 7, 1851 AD." According to details on Schoemann's whereabouts in Indonesia, the place and time of writing are right. This supports the theory that the manuscript was written in response to Schoemann's request.

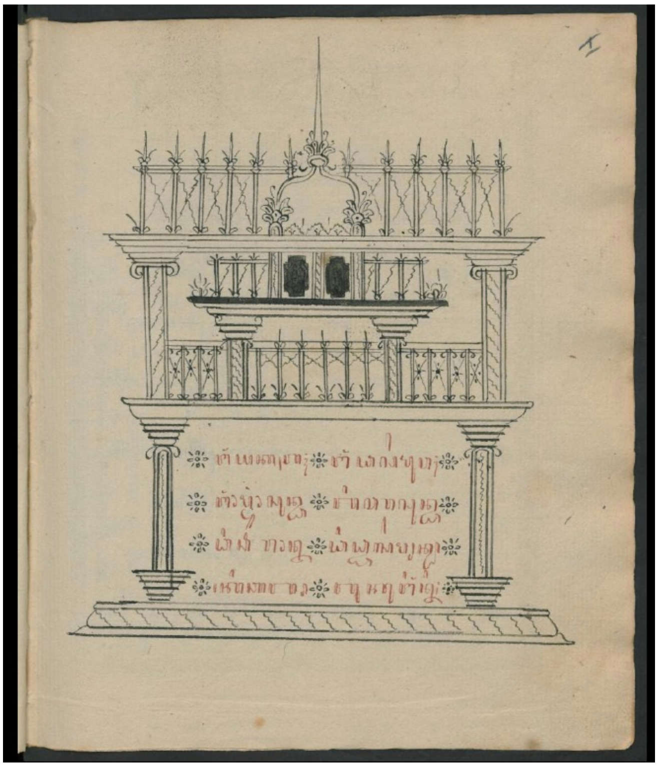


Figure 1. First page of the Schoeman IV 2 manuscript

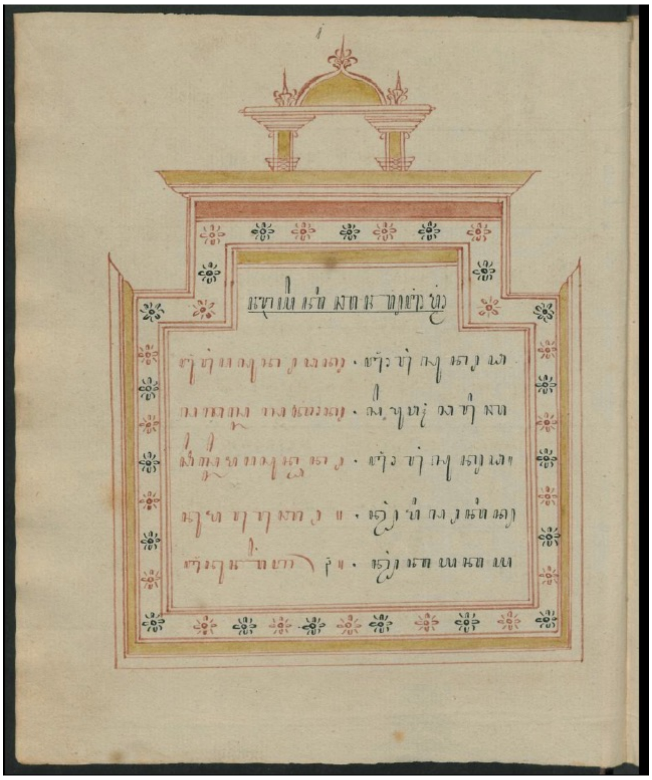


Figure 2. Second page of Schoeman IV 2 manuscript

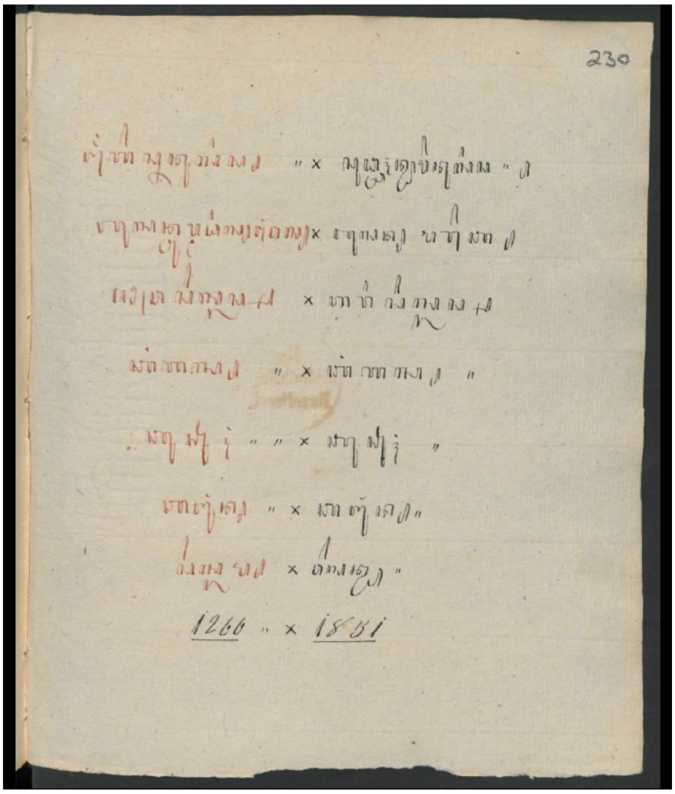

230
1266 × 1851

Figure 3. Final page and colophon of the Schoeman IV 2 manuscript

Lexicon

There isn’t a title for this manuscript. The contents of the manuscript are written in Sundanese and Malay on the title page: “ada semua bicara Sunda didalemnya buku ini.” Several stories in this manuscript contain vocabulary details “Ganti lalakon” or “ganti cerita” is the first sentence in each description. The first story (pages 7–10) is about a fish seller with a daughter and five sons. Raden Kusumah, a talented musician, is the subject of the second story (pages 11-16). These two stories contain Sundanese stories with Malay translations. In Sundanese, there is no special vocabulary in the word list.

The third story is about a blind well digger who collects a large number of fish and sells them to supplement his income (pages 17-26). The fourth story is about a group of Chinese farmers who converted to Islam (pages 27-50). The vocabulary in these two stories is specific to animal lexicons. As shown in table 1, animal lexicons include *munding* (buffalo), *kuda* (horse), *unggas* (poultry), *manuk* (bird), and *lauk* (fish).

Table 1. Animal Lexicon

No.	**Lexicon**	
	Sundanese	Malay
1	munding hideung	kerbo hitam
2	munding bulé	kerbo bulé
3	munding lalaki	kerbo laki-laki
4	munding jalu	kerbo lalaki
5	munding awéwé	kerbo prempuwan
6	kuda bulu beureum	kuda bulu merah
7	kuda waruk rudung	kuda wuk rudung
8	kuda bulu hideng	kuda bulu ngitem
9	kuda bulu bodas	kuda bulu putih

10	kuda bulu héjo	kuda bulu ijo
11	kuda beureum kolot	kuda merah tuwa
12	kuda beureum ngora	kuda merah muda
13	kuda péngkor	kuda pincang
14	hayam diadu	ayam bakalring
15	hayam jago	ayam lalaki
16	hayam jajangkar	ayam lalaki
17	hayam laki	ayam laki-laki
18	hayam bikang	ayam prempuan
19	hayam awéwé	ayam prempuan
20	hayam denten	ayam prempuan
21	hayam turundul	ayam trondol
22	hayam kampu	ayam kecil
23	hayam pitik	ayam kecil
24	hayam leutik	ayam kecil
25	hayam kukut	ayam piaraan
26	hayam jangkung	ayam tinggi
27	hayam cempa	ayam kate
28	hayam urik	ayam burik
29	hayam hideung	ayam item
30	hayam beureum	ayam merah
31	hayam pupuh	ayam pupuh
32	hayam endog	ayam betelor
33	hayam nyilenglem	ayam ngerem
34	hayam megar	ayam netes
35	hayam mangur	ayam beranak
36	meri konéng buluna	bebek kuning bulunya
37	sowang bodas bulunya	angsa putih bulunya
38	mandila hideung buluna	mendila item bulunya
39	éntog kularung	mendilla kelabu
40	japati hideung buluna	dara item bulunya
41	manuk heulang	burung ulung-ulung
42	manuk gagak	burung gagak
43	manuk tikukur	burung tekukur
44	manuk titiran	burung prekutut
45	manuk piit	burung perit
46	manuk galatik	burung gelatik
47	manuk kakatuwa	burung kakatuwa
48	manuk séréndét	burung seréndet
49	manuk nuri	burung nuri
50	manuk kérak	burung kaléng
51	manuk garéja	burung geréja
52	manuk kapinis	burung lewari
53	manuk merak	burung merak
54	manuk bango	burung bango
55	manuk kuntul	burung kuntul
56	manuk walilis	burung walilis
57	manuk ngapung	burung ngaspur
58	manuk cangkurileng	burung kutilang
59	manuk ciwung	burung beyo
60	manuk saéran	burung kucica
61	manuk ékék	burung bétét
62	lauk beunteur	ikan banter
63	lauk ngurang	ikan nguda(ng)
64	lauk sénggal	ikan senggal
65	lauk géde	ikan besar

66	lauk cai	ikan kali
67	lauk deleg	ikan gabus
68	lauk lélé	ikan lélé
69	lauk bogo	ikan bogo
70	lauk nilem	ikan nilem
71	lauk jambal	ikan jambal
72	lauk lika	ikan likla
73	lauk bangu	ikan bangu
74	lauk kancra	ikan kanjra
75	lauk léyat	ikan léyat
76	lauk beureum mata	ikan merah soca
77	lauk hampala	ikan ampala
78	lauk bawal	ikan bawal
79	lauk génggéhak	ikan génggéhak
80	lauk sénggal	ikan sénggal
81	lauk berod	ikan berod
82	lauk benter	ikan benter
83	lauk hurang	ikan udang
84	lauk emas	ikan emas
85	lauk guramé	ikan guramé
86	lauk kancra emas	ikan kancra emas
87	lauk paray	ikan paray
88	lauk wadon	ikan wadon
89	lauk pangsét	ikan ngasin
90	lauk pépéték	ikan pépéték
91	lauk peda	ikan peda
92	lauk tuhur	ikan kering
93	lauk teri	ikan teri
94	lauk sepat	ikan sepat
95	lauk telang-talang	ikan talang-talang
96	lank témjang	ikan témbang
97	lauk cécéré	ikan cécéré
98	lauk selar koneng	ikan selar kuning
99	lauk mata kucing	ikan mata kucing
100	lauk cumi-cumi	ikan cumi-cumi
101	lauk betok	ikan betok
102	lauk tanggiri	ikan tenggiri
103	lauk totongkol	ikan totongkol
104	lauk cucut	ikan cucut
105	lauk layur	ikan layur
106	lauk nun	ikan nun
107	lauk julung-julung	ikan julung-julung
108	lauk lodan	ikan ludan
109	lauk keting	ikan keting
110	lauk pari	ikan pari
111	lauk lempuk	ikan lempuk
112	lauk sabelah	ikan sabelah
113	lauk soro	ikan soro
114	lauk poro jontor	ikan tambrara
115	lauk lendi	ikan lendi
116	lauk biru	ikan biru
117	lauk manyung	ikan manyung
118	lauk méngnga	ikan méngnga
119	lauk kembung	ikan kembung
120	lauk kuya	ikan bulus
121	lauk bayawak	ikan meyawak

122	lauk buhaya	ikan buwaya
123	lauk penyu	ikan penyu
124	lauk keyep	ikan kapiting
125	lauk kura-kura	ikan kura-kura
126	lauk rajungan	ikan rajungan
127	lauk kerang	ikan kerang
128	lauk remis	ikan remis
129	lauk uncal	ikan menjangan
130	lauk mencek	ikan kijang
131	lauk peucang	ikan kancil
132	lauk kelenci	ikan kelinci
133	lauk pesing	ikan tenggiling
134	lauk oray	ikan ular
135	lauk léncah	ikan lencah
136	lauk ajag	ikan ajag
137	lauk bagong	ikan babi

Unfortunately, the text jumps to page 125 after page 50. Pages 125 to 158 have been completed. The previous story's narration is cut off. As seen in table 2, the text explicitly includes the lexicon of household appliances from page 125 to page 128.

Table 2. Home appliances lexicon

No.	Lexicon	
	Sundanese	Malay
1	pakakas	perabot
2	piring mangkok	piring mangkok
3	padéhan	tangkepan
4	rampadan	dulang-dulang
5	aseupan	kukusan
6	hawu	dapur
7	pawon	dapur
8	suluh	kayu
9	cowét	piring tanah
10	boboko	bakul
11	sangid	bakul
12	nyiru	tetampah
13	hihid	kipas
14	téssi	pénjak
15	comlong	mangkok
16	wadah uyah	tempat garem
17	para pawon	loténg dapur
18	téko	kétél
19	tikuk	kétél
20	cécémpeh	tatampah
21	pabéyasan	paberasan
22	siwurna	gayungnyah
23	sinduk angen	séndok sayur
24	songsongna	semperong

The text includes lexicons linked to cooking activities on pages 128 to 132. Table 3 lists eleven lexicons for cooking activities.

Table 3. Cooking activity lexicon

No.	Lexicon	
	Sundanese	Malay
1	geura ngéjo	lekas masak
2	geura ngangeun	lekas nyayur
3	geura nyambel	lekas nyambel
4	geura ngisikan	lekas cuci beras
5	geura ngumbah béyas	lekas cuci beras
6	geura ngaliwet	lekas masak
7	ngabubur	lekas ngalimpa
8	geura nutuwan	lekas numbuk
9	geura ngasakan	lekas matengin
10	masing ngagolak	biyar mendidih
11	hurungkeun seuneu	manyalain apinya

Page 133 begins with the phrase "ganti deui carita" or the story is changed again. As shown in Table 4, pages 133 to 138 contain 49 vocabularies related to soil types or conditions.

Table 4. Lexicon of soil types or conditions

No.	lexicon	
	Sundanese	Malay
1	taneuh luhur	tanah tinggi
2	taneuh handap	tanah randa(h)
3	taneuh tegal	tanah padang
4	taneuh lamping	tanah jurang
5	taneuh lombang	tanah kobak
6	tanah gawir	tanah jurang
7	taneuh ipis	tanah tipis
8	taneuh lémpar	tanah rata
9	taneuh biyé	tanah leyur
10	taneuh dempak	tanah cepak
11	taneuh bunder	tanah bulat
12	taneuh buled	tanah bulat
13	taneuh lingih	tanah licin
14	taneuh leueur	tanah licin
15	taneuh kéros	tanah kerus
16	taneuh monyong	tanah terup
17	taneuh nyungcu	tanah terus
18	taneuh nérop	tanah terus
19	taneuh lega	tanah lébar
20	taneuh rubak	tanah lébar
21	tanah jauh	tanah jaoh
22	taneuh panjang	tanah panjang
23	taneuh lawas	tanah lama
24	taneuh lila	tanah lama
25	taneuh kakara	tanah baru
26	taneuh anyar	tanah baru
27	taneuh cikénéh	tanah barusan
28	taneuh biye	tanah baru
29	taneuh baréto	tanah dahulu
30	taneuh kimpel	tanah kempel
31	taneuh badag	tanah besar
32	taneuh leles	tanah lemes
33	taneuh lemer	tanah tepis

34	taneuh kandel	tanah tebel
35	taneuh bau	tanah busuk
36	taneuh nyonyos	tanah busuk
37	taneuh buruk	tanah busuk
38	taneuh pait	tanah pangit
39	taneuh amis	tanah manis
40	taneuh pangsét	tanah asin
41	taneuh tuur (tuhur/tuus)	tanah kering
42	taneuh garing	tanah kering
43	taneuh kélénténg	tanah kering
44	taneuh ibul	tanah isep
45	taneuh lebu	tanah abu
46	taneuh beureum	tanah mérah
47	taneuh bodas	tanah putih
48	taneuh hideung	tanah item
49	tanah héjo	tanah ijo

Page 139 begins with the phrase “lalakon imah,” which means “tale of the home.” As shown in Table 5, pages 139 to 144 contain 50 vocabulary words related to the types or conditions of the property.

Table 5. Lexicon of types or house conditions

No.	Lexicon	
	Sundanese	Malay
1	imah luhur	rumah tinggi
2	imah handap	rumah rendah
3	imah ranggon	rumah panggung
4	imah joglo	rumah lelimasan
5	imah témbok	rumah témbok
6	imah papan	rumah papan
7	imah gebyog	rumah papan
8	imah awi	rumah bangbu
9	imah kai	rumah kayu
10	imah buruk	rumah busuk
11	imah weuteuh	rumah bahru
12	imah anyar kénéh	rumah bahru abes
13	imah kara anggeus	rumah baru abes
14	imah kara adeg	rumah baru bediri
15	imah mesak	rumah bagus
16	imah kenténg	rumah kenténg
17	imah dijiyeun	rumah dibikin
18	imah butut	rumah copong
19	imah warang	rumah jarang
20	imah rawing	rumah rawék
21	imah rangsak	rumah bolong
22	imah rajét	rumah bolong
23	imah kancang	rumah jarang
24	imah caang	rumah terang
25	imah rubat	rumah busuk
26	imah rabét	rumah busuk
27	imah goréng	rumah jelék
28	imah anggang	rumah renggang
29	imah logor	rumah longgar
30	imah loncér	rumah longgar
31	imah leueur	rumah licin

32	imah ungcutan	rumah telucut
33	imah runtuh	rumah rubuh
34	imah rubuh	rumah rubuh
35	imah rugrug	rumah jatoh
36	imah murag	rumah jatoh
37	imah ticongkél	rumah colok
38	imah rugrug	rumah rubuh
39	imah rentas	rumah patah
40	imah penggas	rumah patah
41	imah potong	rumah patah
42	imah tijungkel	rumah songlot
43	imah ropoh	rumah butut
44	imah uduh	rumah amoh
45	imah ahéng	rumah bagus
46	imah kalép	rumah bagus
47	imah bobo	rumah amoh
48	imah emoy	rumah amoh
49	gedogan	istal
50	kandang	kandang

The sentence “lain lalakon” or another story begins on page 145. As shown in Table 6, several lexicons of verbs and adjectives can be found on pages 145 to 155.

Table 6. Lexicon of verbs and adjectives

No.	Lexicon	
	Sundanese	Malay
1	cokot	ambil
2	béré	kasih
3	biken	kasih
4	okod	ambil
5	sangeuk	terada mau
6	narimakeun	tarimaan
7	ngagebeg	terkejut
8	teundeun	taro
9	reuwas	terkejut
10	ngarénjag	terkejut
11	ngaranjug	terkejut
12	suker	susah
13	susah	susah
14	teu gaduh	terada punya
15	séwot	marah
16	ambek	marah
17	ingsrek	ngisap
18	nyesep	ngisap
19	kokoro	miskin
20	ngadéngé	mendenger
21	moncor	keluwar
22	bijil	keluwar
23	asup	masuk
24	hadé	baék
25	teu hadé	tiada baék
26	goréng	jelék
27	teu goréng	tida jelék
28	disimbut	disimbut

29	simbutan	selimutan
30	hareudang	gerah
31	hanteu hareudang	tiada gerah
32	tiris	dingin
33	teu tiris	tiada dingin
34	panas	panas
35	teu panas	tiada panas
36	ingin	ingin
37	teu ingin	tiada ingin
38	kerong	tusuk
39	kerongan	tusukin
40	obah	berobah
41	henteu obah	tiada berobah
42	cicing	diem
43	cicingkeun	diemin
44	sebentar	sebentar
45	ngahinghing	merintih

Several lexicons of designations or kinship links can be found on pages 156 to 158, 194, 196, and 197. Table 7 displays the lexicon.

Table 7. Lexicon of designations or kinship relations

No.	Lexicon	
	Sundanese	Malay
1	awéwé	prempuwan
2	lalaki	laki-laki
3	randa	randa
4	lanjang	perawan
5	cawénné	perawan
6	kolot	tuwa
7	budak	anak kecil
8	aki-aki	kaki-kaki
9	nini-nini	néné-néné
10	parawan	perawan
11	lanjang	perawan
12	ngora	muda

Conclusion

The Schoemann IV 2 manuscript is a compilation manuscript from the *Staatsbibliothek zu Berlin* that contains several Sundanese lexicons with Malay translations. The text contains 137 animal lexicons, 24 lexicons of household utensils, 11 lexicons of cooking practices, 49 lexicons of types or soil conditions, 50 lexicons of types or house conditions, 45 lexicons of verbs and adjectives, and 12 lexicons of classification or kinship written in *cacarakan* script. To uncover a more complete history of Indonesian dictionaries, research on lexicographic works written by Indonesians themselves is needed.

References

Ali, Lukman (2000). *Sejarah Ejaan Bahasa Indonesia*. Jakarta: Pusat Bahasa.

Chambert-Loir, Henry dan Oman Fathurahman. (1999). *Khazanah Naskah: Panduan Koleksi Naskah-Naskah Indonesia Sedunia.* Jakarta: Yayasan Obor Indonesia.

Coolsma, S. (1985). *Tata Bahasa Sunda*. Jakarta: Djambatan.

Hidayat, Rahayu Surtiati. (2018). *Hakikat Ilmu Pengetahuan Budaya*. Jakarta: Yayasan Pustaka Obor Indonesia.

Linehan, W. (1949). The Earliest Word-lists and Dictionaries of the Malay Language. *Journal of the Malayan Branch of the Royal Asiatic Society*, 22 (1), 183-187.

Proudfoot, I. (1993) *Early Malay Printed Books. A Provisional Account of Materials Published in the Singapore-Malaysia Area up to 1920, Noting Holdings in Major Public Collections*. Academy of Malay Studies and The Library, University of Malaya.

Tim Penyusun. (1995). *Kamus Besar Bahasa Indonesia Edisi Kedua*. Jakarta: Balai Pustaka.

ELECTRONIC BILINGUAL DICTIONARY FOR LEARNING ENGLISH AT JUNIOR HIGH SCHOOL

Asti Ramadhani Endah Lestari[1], Nuruddin[2], Yumna Rasyid[2]
[1]Universitas Indraprasta PGRI, Indonesia; [2]Universitas Negeri Jakarta, Indonesia
asti.ramadhani@yahoo.co.id

Abstract

This article is the continuation of our previous research article entitled "The Preferences and Needs of Electronic Dictionary among Junior High Students in Jakarta. Dictionaries and learning English as a foreign language are two inseparable things. Following the development of technology, students have begun to leave printed dictionaries and switch to electronic dictionaries. The available electronic bilingual dictionaries in Indonesia are not specifically designed to be used in the learning process in the English classroom. Therefore, it is necessary to review the suitability and appropriacy between electronic bilingual dictionaries and the needs of students and teachers as the users. This article will describe the results of the analysis of microstructures and macrostructures of electronic dictionaries that are often used by Junior High School students and compared the results with the needs of students and teachers as the users. There are two electronic bilingual dictionaries analyzed in this research, namely kamusku and google translate. The dictionaries were chosen after the result of the survey stated that Junior High School students use these two dictionaries. The result of this research surprisingly concluded that the available electronic bilingual dictionaries do not meet the needs of the user and it is inappropriate to be used in learning English at Junior High School Level. This inappropriacy affects the effectiveness of the dictionary because this condition makes students feel confuses when using the dictionary. The result of the research figured out that Junior High School students need a dictionary that is specifically designed for the purpose of learning English suitable with their level.

Keywords: Electronic dictionary, bilingual dictionary, learning English, Junior High School

1 Introduction

The dictionary is one of the learning media that can help implement the learning process of English as a foreign language. A dictionary cannot be separated from the person who is studying a foreign language. Whenever someone wanted to know the meaning of a word, he would look it up in the dictionary. The use of dictionaries in the learning process can improve student's vocabulary skills. This statement is supported by Asgari & Mustapha (2011), which is that dictionaries are excellent learning media in improving vocabulary. Even so, Nesi & Bae (2014) say that there are not many studies discussing dictionaries because dictionaries are an individual learning medium even though the dictionary is essential in learning foreign languages. Therefore, the researchers feel that the dictionary is an excellent topic to be discussed in this research.

The use of electronic dictionaries has dominated nowadays. This phenomenon is supported by the results of interviews with 2 English teachers at two different schools who said that dictionaries have a vital role in the learning process. The teacher stated that students prefer to use an electronic dictionary even though they advised students to use a printed dictionary. The result is in line with research conducted by (Alhaisoni, 2016; Ebanéga & Moussavou, 2008) that most students prefer to use a bilingual electronic dictionary rather than a printed or monolingual electronic dictionary.

Another research conducted by (Şevik, 2014)regarded as lexicographical reference books, are considered as indispensable learning tools in foreign language acquisition. It seems that the recent advances in IT change and shape EFL learners' dictionary ownership and preferences. Research on EFL learners' dictionary ownership and preferences has been increasing in abroad EFL contexts to explore this new situation especially over the past decade. Such research mainly result that paper dictionaries are losing popularity and that electronic dictionaries are gaining importance among EFL learners (e.g. Jian et al., 2009 and Kobayashi, 2008 that 96.82% of students had a printed dictionary, 92.2% of students had a dictionary in the form of a mobile phone application, 57.96% of students had a dictionary on a laptop / PC, and only 28% of students had a dictionary in the form of a mobile phone application. Meanwhile, the results of research related to the use of dictionaries showed that 92.2% of students used a dictionary in the form of a mobile phone application, 84.71% of students used a printed dictionary, 57.96% of students used a dictionary on a laptop / PC, and 28% of students used a pocket dictionary. By looking at the results of this study, we can conclude that all students who have electronic dictionaries in the form of applications for mobile phones, laptops, PCs, or pocket dictionaries use the dictionary. Meanwhile, some students who have printed dictionaries do not use them. Barham (2017) also shows that students also use electronic dictionaries in other subjects outside of the classroom and when teaching or helping their siblings who are studying.

Zheng & Wang (2016) define an electronic dictionary as a portable electronic device that functions as a digital form of any dictionary. The electronic dictionary consists of several forms such as CD-ROM, DVD-ROM, network, or applications on mobile phones. The electronic dictionary also has various functions ranging from a general monolingual dictionary, a bilingual general dictionary, a student dictionary, a terminology dictionary, a thesaurus, etc. Granger (2012) states that technology integration in the dictionary brings lexicography in a better direction. Along with technological developments, dictionaries are limited to printed form and electronic form ranging from dictionaries that can be accessed online or offline.

The results of previous research conducted by Rezaei & Davoudi (2016) found that the most effective media for a dictionary is electronic media. Other research conducted by Alharbi, (2016) and Rezaei & Davoudi (2016) also shows that electronic dictionaries provide more benefits than printed dictionaries in improving student's vocabulary skills. Students who use electronic dictionaries score higher than students who use printed dictionaries. The difference in the effectiveness of this dictionary is because electronic dictionaries have different features from printed dictionaries. Previous research conducted by Chan (2017) found that when using a dictionary, students did four things, namely identifying the meaning of the target language in their mother tongue, looking at examples of word use in sentences, looking at the syntactic structure of the word, and trying to enter the language word. English into the syntactic structure listed in the dictionary. Using the dictionary by the target user can help the lexicographers determine the suitable microstructure and macrostructure for the dictionary.

Generally, Jackson (2013) said that the dictionary consists of two parts, namely macrostructure and microstructure. The macrostructure consists of three parts: the front matter, the body, and the appendices, while the microstructure refers to the arrangement of information contained in the dictionary. The introduction section contains a foreword that explains the advantages of the dictionary or the revisions made in the previous section. In addition, the introduction also contains other information such as instructions for using the dictionary, the transcription system of speech used, and a list of abbreviations. Meanwhile, the core part contains a list of words followed by supporting information. It is this core part that intersects with the microstructure in the dictionary. The microstructure of a dictionary differs depending on its purpose. However, in general, the microstructure of a dictionary consists of spelling, pronunciation, affixes, word classes, definitions, examples of usage, and etymology.

Tan & Woods (2008) describe three features of an electronic dictionary: macro-structural features, microstructural features, and inter-structural and mediostuctural features. Whereas the macrostructure in a print dictionary refers to the composing features, the macrostructure refers to how users can access entries in an electronic dictionary. In electronic dictionaries, the order of the entry is not an important consideration. Microstructure refers to the comprehensive information about an entry that is contained in the dictionary. Finally, the interstructural and mediostructural features refer to how an entry in the dictionary can be integrated with other sources outside the dictionary.

Heuberger (2016) explains that electronic dictionaries have three other advantages compared to printed dictionaries, namely customization, hybridization, and user input. The first aspect is the adjustment. The purpose of customization as the advantage of an electronic dictionary is that an electronic dictionary can be adapted according to user needs. The second aspect is hybridization. The differences between dictionaries, encyclopaedias, databases, and translation tools are not significant in electronic media. Meanwhile, the concept of user input refers to the dictionary feature that allows users to enter entry or gloss.

Oppentocht & Schutz (2003) describe three main advantages of electronic dictionaries over traditional dictionaries: easier access and more explicit information, improved dictionary functions, and dictionary extensions. In more detail, the advantages of electronic dictionaries will be explained below:

a. easier access and more explicit information

In the printed dictionary, users find many abbreviations that sometimes confuse users, such as adj, bre, adv, inf, etc. The electronic dictionary does not limit the number of words in the dictionary, so the lexicographer does not need to make these abbreviations. In a traditional dictionary, users have to open several pages to cross-reference, which is usually marked with the word (see). The hyperlink feature in the electronic dictionary enables users to cross-reference without having to type the word when searching a new word.

b. improved dictionary function

Print dictionary users need special skills in order to use the dictionary quickly and effectively. For example, the user must understand the order of the dictionaries, whether by theme or alphabetically. Electronic dictionary users do not need special skills; just by typing the desired word, then the user can obtain the word's definition. Another advantage of the electronic dictionary is the voice feature. Users can hear the pronunciation of a word correctly without reading the phonetic letters like those in a printed dictionary.

c. dictionary extension

There is a hyperlink feature in the electronic dictionary that can help users find examples of the use of words in sentences in other sources. Nowadays, many computer programs or internet pages have translation features integrated with an electronic dictionary.

The effectiveness of a dictionary as a learning medium is primarily determined by the characteristics of the dictionary user. Cimermanová (2012) argues that the criteria for a good dictionary are mainly determined by user needs, whether the user uses a dictionary to learn a science or to be able to understand a language while traveling. The completeness of entries, defining vocabulary, and other information that needs to be included in the dictionary must be tailored to the user's needs. Therefore, a lexicographer must target the dictionary user specifically.

Kwary (2018) supported this opinion by saying that a dictionary created without specifying clear user criteria has no benefit at all. Therefore, determining user criteria and user requirements is very important. Therefore, Nesi (2013) states that lexicographic research should focus more on specific dictionaries that can be used in certain groups.

Atkins & Rundell (2008) divide student dictionaries into three types: monolingual dictionaries, bilingual dictionaries, and multilingual dictionaries. A monolingual dictionary is a dictionary that contains only one language, such as the Kamus Besar Bahasa Indonesia. A bilingual dictionary is a dictionary that consists of two languages, such as the English-Indonesian dictionary. Meanwhile, a multilingual dictionary consists of three or more languages, such as the English-Indonesian-Arabic dictionary. The bilingual dictionary is divided into two, namely unidirectional and bidirectional. A unidirectional dictionary is a dictionary that contains words from a language and is translated into the target language, for example, the English-Indonesian dictionary. Unidirectional dictionaries are divided into active dictionaries (encoding dictionary) and passive dictionaries (decoding dictionary). An active dictionary is a dictionary whose source language is the user's native language. On the other hand, a passive dictionary is a dictionary that makes the mother tongue the target language. Meanwhile, a bidirectional dictionary is a dictionary that consists of two parts, for example, a part containing words from English and translated into Indonesian and a part containing words from Indonesian and translated into English.

Nkomo & Madiba (2012) explain that a bilingual dictionary has two functions, namely as a tool in translating and also as a reference in a monolingual dictionary. When students use a monolingual dictionary and find some defining vocabulary words that are not understood, students still have to look for the meaning of the word in the bilingual dictionary. Golavar, Beikian, Nooramin, & Firoozkoohi (2012) found that a practical dictionary for beginner-level foreign language learners is a bilingual dictionary, a dictionary consisting of two languages. In this study, the bilingual dictionary refers to the English-Indonesian dictionary. The bilingual dictionary helps foreign language learners find the equivalent word between their native language and their learning. The dictionary helps students understand the vocabulary of the topic being studied and helps students choose the correct vocabulary to use in the text to be created.

A complete dictionary such as Kamus Besar Bahasa Indonesia (KBBI) can even be said to be not ideal if used by the beginner level of Bahasa Indonesia as a foreign language. The characteristic of different dictionary users requires a different type of dictionary. Kwary (2003) states that errors in determining the characteristics of target users can affect the effectiveness of the dictionary. One of the characteristics of dictionary users is the user's age. Each different age range requires a different type of dictionary and information. Children tend to need different information from adults (Tono, 2012). This statement is supported by previous research conducted by Nkomo & Madiba (2012), which shows that a good student dictionary uses a limited corpus following the characteristics of students as target users. A dictionary made with a specialized corpus helps students understand vocabulary faster. The words contained in the dictionary are the words that are used mainly by students.

Wild, Kilgarriff, & Tugwell (2013) research shows that the corpus of children is different from the general corpus. The corpus of children should be different from the adult corpus because the text used by children is different from the adult text. This study shows that the child's corpus has many benefits for learning, namely to determine the list of entries in the child's dictionary, to identify collocations and keywords that are often used by children, to provide examples of the use of appropriate words for children, and to keep abreast of changes. in children's language.

As a learning media, student dictionaries must be in line with the current curriculum. Each level of education requires a different student dictionary. In implementing the 2013 curriculum in Indonesia, formal English learning begins at the junior high school level. The dictionary circulating in Indonesia is still a general dictionary. There is no single English-Indonesian bilingual dictionary specifically designed to learn English at the specific unit level, especially in junior high schools. This condition will affect the level of effectiveness of a dictionary in the learning process.

Based on the phenomena and theoretical background explained above, the researcher feels the need to conduct research that focuses on discussing the micro and macrostructure of the electronic dictionary used by students. The results of this analysis are expected to provide significant input to determine whether the electronic dictionary commonly used by students is effective in helping improve vocabulary understanding in English learning for junior high school students in Indonesia.

2 Method

This research uses a qualitative approach with a comparative descriptive method. The instrument used in this study was observation. The researcher observed the micro and macrostructure of each dictionary. Researchers analyzed the dictionary using an observation guide made based on the Jackson (2013) theory. Then, the researcher compared the findings of each dictionary to see the effectiveness of the dictionary in terms of foreign language learning.

3 Result

In this study, researchers analyzed the microstructure and macrostructure of two dictionaries. The researcher chose the two dictionaries that students most widely used. Researchers conducted preliminary research and found that many students use kamusku and google translate.

a) Kamusku

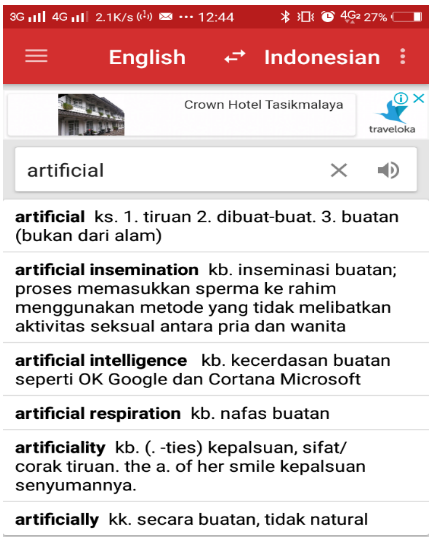
Kamusku is an offline dictionary application made by the Kodelokus company. The user can download Kamusku for free from the google play store. This application also provides a premium or paid version, but all students use the standard version of the application so that researchers only analyze the standard version of the kamusku application. The following is a description of the analysis results of the macrostructure and microstructure from kamusku.

1) macrostructure

There are 6 features available in Kamusku. First, Kamusku has two versions, namely English-Indonesian and Indonesian-English. Second is Bookmark. The user uses the bookmark feature to mark favourite words. Third is History. This feature allows the user to recall words. This feature helps users so that users do not have to retype words that have been searched through the Kamusku application. Fourth is About. This feature contains information about Kodelokus, the company that made this application. Fifth is Google speech recognition. This feature allows users not to type words when searching the words manually. Sixth is Examples of pronunciation. This feature allows users to hear the correct pronunciation of the word.

2) Microstructure

We can see the microstructure of Kamusku in the picture below.



artificial

artificial ks. 1. tiruan 2. dibuat-buat. 3. buatan (bukan dari alam)

artificial insemination kb. inseminasi buatan; proses memasukkan sperma ke rahim menggunakan metode yang tidak melibatkan aktivitas seksual antara pria dan wanita

artificial intelligence kb. kecerdasan buatan seperti OK Google dan Cortana Microsoft

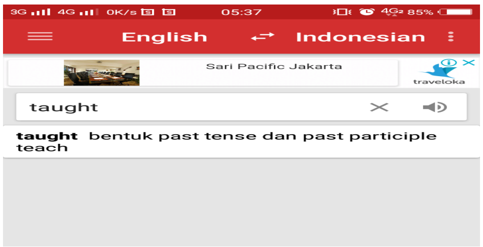
artificial respiration kb. nafas buatan

artificiality kb. (. -ties) kepalsuan, sifat/corak tiruan. the a. of her smile kepalsuan senyumannya.

artificially kk. secara buatan, tidak natural

Figure 1 microstructure of Kamusku

The app designer uses bold to mark entries in Kamusku. Kamusku contains information about word classes. The markers used for word classes are "kkt" for transitive verbs, "kki" for intransitive verbs, "kb" for nouns and "ks" for adjectives. The past tense of a verb, both regular and irregular verbs, is considered an entry. The past tense of regular verbs is denoted by the symbol "kkt" or a transitive verb and has the meaning of the passive form—example of the word cooked as shown below. Kamusku indicates the past tense of irregular verbs as an individual entry.

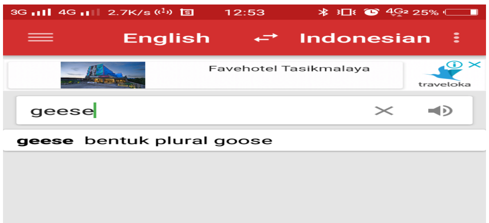


taught

taught bentuk past tense dan past participle teach

Figure 2 The display of irregular verbs in Kamusku

The plural form of irregular noun is considered as an entry.

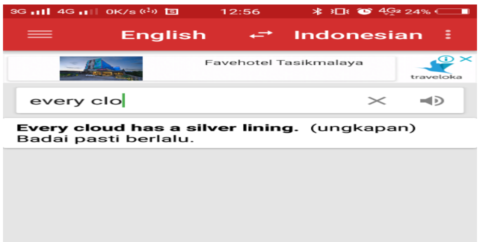


geese

geese bentuk plural goose

Figure 3 The display of irregular noun in Kamusku

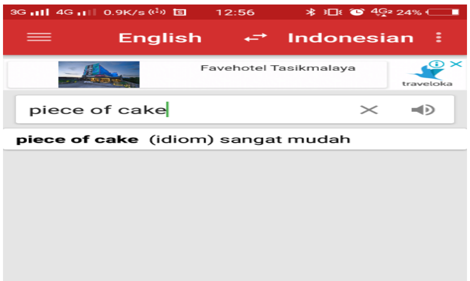
Kamusku also indicates idiom and quotes as one individual entry as seen in picture below.



every clo

Every cloud has a silver lining. (ungkapan) Badai pasti berlalu.

Figure 4 The display of quotes in Kamusku



piece of cake

piece of cake (idiom) sangat mudah

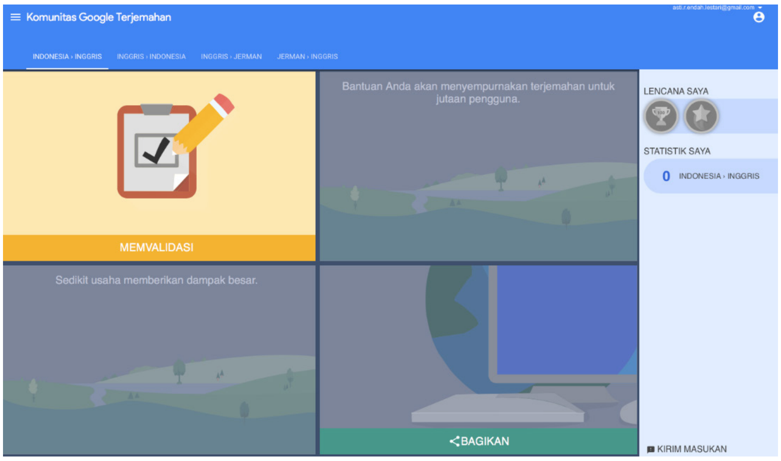
Figure 5 The display of idiom in Kamusku

b) Google Translate

The second dictionary analyzed is google translate. This dictionary is an online dictionary created by google. Google translate not only contains English-Indonesian and Indonesian-English dictionaries but also contains dictionaries from various other languages. The following will describe the analysis of the macrostructure and microstructure from Google Translate.

1) macrostructure

Google Translate has two versions, namely the English-Indonesian and Indonesian-English versions. Second is History. This feature allows the user to recall words. This feature helps users so that users do not have to retype words that have been searched. Third is Bookmark. The user uses the bookmark feature to mark favourite words. Fourth is Community. In this feature, users can validate or correct translations made by Google, as shown in the image below.

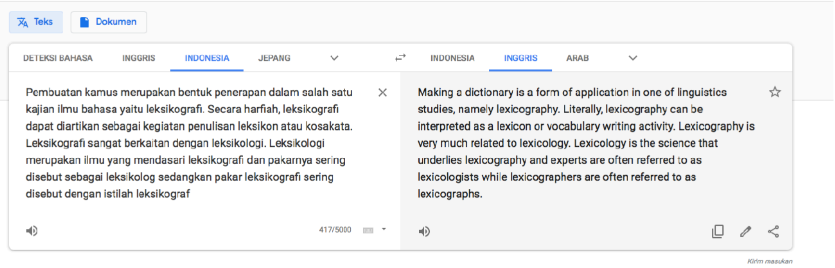


Bantuan Anda akan menyempurnakan terjemahan untuk jutaan pengguna.

Sedikit usaha memberikan dampak besar.

Figure 6 The display of community feature in google translate

Fifth is Google speech recognition. This feature is only available in the mobile version of Google Translate. This feature uses the Google Speech Recognition application, where users can search for the meaning of a word or sentence by saying the word. This feature allows users not to type words when searching the words manually. Sixth is Sentence translation. Google Translate has a feature that can translate sentences and even text up to 5,000 words. However, this feature is still ineffective because many translation errors are made in this feature, especially if the user wants to translate from Indonesian to English. Like the example below.



Pembuatan kamus merupakan bentuk penerapan dalam salah satu kajian ilmu bahasa yaitu leksikografi. Secara harfiah, leksikografi dapat diartikan sebagai kegiatan penulisan leksikon atau kosakata. Leksikografi sangat berkaitan dengan leksikologi. Leksikologi merupakan ilmu yang mendasari leksikografi dan pakarnya sering disebut sebagai leksikolog sedangkan pakar leksikografi sering disebut dengan istilah leksikograf

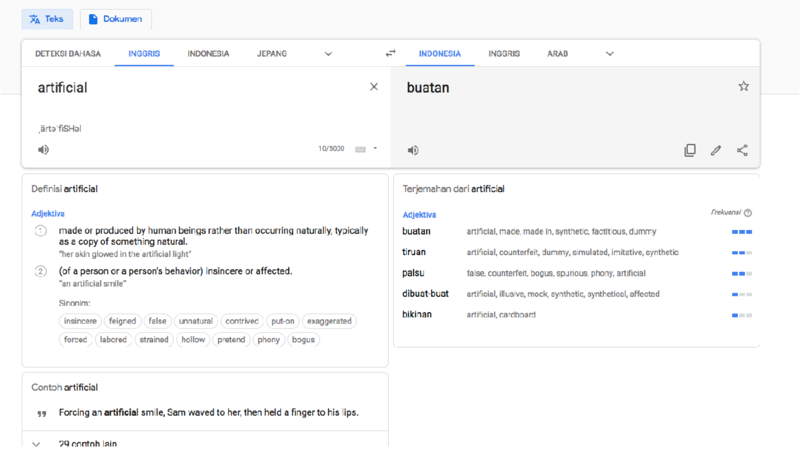
Making a dictionary is a form of application in one of linguistics studies, namely lexicography. Literally, lexicography can be interpreted as a lexicon or vocabulary writing activity. Lexicography is very much related to lexicology. Lexicology is the science that underlies lexicography and experts are often referred to as lexicologists while lexicographers are often referred to as lexicographs.

Figure 7 The display of sentence translation by using Google Translate

Seventh is Sentence translation. Google translate has a feature that can translate sentences and even text up to 5,000 words. However, this feature is still ineffective because many translation errors are made in this feature, especially if the user wants to translate from Indonesian to English. Like the example below. And Finally is Suggest edit. Google translate allows users to be able to provide suggestions for improvements to the translated text.

2) microstructure

Microstructure of Google Translate can be seen in the following picture.



artificial

ˌärtəˈfiSHəl

buatan

Definisi **artificial**

Adjektiva

1. made or produced by human beings rather than occurring naturally, typically as a copy of something natural.
"her skin glowed in the artificial light"
2. (of a person or a person's behavior) insincere or affected.
"an artificial smile"

Sinonim:
insincere feigned false unnatural contrived put-on exaggerated forced labored strained hollow pretend phony bogus

Terjemahan dari **artificial**

Adjektiva		Frekuensi
buatan	artificial, made, made in, synthetic, factitious, dummy	■■■
tiruan	artificial, counterfeit, dummy, simulated, imitative, synthetic	■■□
palsu	false, counterfeit, bogus, spurious, phony, artificial	■■□
dibuat-buat	artificial, illusive, mock, synthetic, synthetical, affected	■□□
bikinan	artificial, cardboard	■□□

Contoh **artificial**

Forcing an **artificial** smile, Sam waved to her, then held a finger to his lips.

29 contoh lain

Figure 8 The display of microstructure of Google Translate

Google Translate displays entry and gloss in two different columns. Information about word classes is written explicitly, such as nouns, verbs, adjectives, adverbs, prepositions, and conjunctions. There is information about the frequency of use of words in the corpus. In some entries, there is information about word synonyms. There is information about examples of how to use words in sentence.

4 Analysis and Discussion

The dictionary is a learning media that is inseparable from foreign language learning. The dictionary can improve student's language skills, especially in terms of vocabulary mastery. Research conducted by Alhaisoni (2016) found that dictionaries are an effective learning medium to improve student's vocabulary skills significantly. With the help of a dictionary, students can understand information about vocabulary very quickly.

We can see from the above analysis results that Kamusku and google translate have similarities and differences. The two dictionaries have several macrostructures in common. The following is what the two dictionaries have in common. First, both dictionaries allow the user to use either encoding and decoding dictionaries. This option helps students improve their receptive skills and productive skills. Encoding dictionary helps students produce a word from the mother tongue to the target language, while the decoding dictionary helps students understand a word from the target language to the mother tongue. Both of these dictionaries also consists of a history feature that helps students recall words searched by students. This feature can make the dictionary more efficient.

Both dictionaries also consist of pronunciation features. This feature helps students to learn how to use a word appropriately (Chaer, 2007). Another advantage that the two dictionaries have is the speech recognition feature that allows students to look for a word without typing it. This feature is a feature created by Google and can be integrated with applications available on the Google Play Store. In the foreign language classroom, this feature can help students improve pronunciation. Students can try to say a word. If the dictionary shows the correct word, then the student has pronounced the word correctly.

In terms of macrostructure, there is one feature available in Kamusku, but it is not available in Google Translate. On the other hand, two features are available in Google Translate but not available in Kamusku, namely the community and sentence translation features. Kamusku is a dictionary that can be accessed offline, so students only need one-time access to the internet. Meanwhile, Google Translate can only be accessed online so that if the students do not connect to the internet, the student cannot use the dictionary.

The community feature allows users to validate words created by Google. In this feature, users can provide suggestions if they find an error from the translation made by Google. Previous research conducted by Wenner & Sköldberg (2019) shows that the user input feature has its appeal to users. However, this feature does not provide significant benefits in English classrooms at the junior high school level.

Another feature that is available in Google Translate but not available in Kamusku is sentence translation. This feature helps users translate text faster. However, this has received criticism regarding grammatical and diction inaccuracy Nugraha (2020) in translating English into Indonesian. In addition, from a learning perspective, this feature can hinder the learning process. If students use this feature too often, students do not need to understand every vocabulary correctly.

In general, these two dictionaries consist of several microstructures in common. However, these two dictionaries display the information differently. Kamusku uses bold print to mark entries, and gloss is arranged horizontally. On the contrary, Google Translate arrange gloss vertically with additional information of synonym. The word class information in Kamusku is written with abbreviations, while on Google Translate, the word class is written explicitly using a color marker. Gloss in Google Translate is also arranged based on the frequency. The frequency of words here is taken from the general corpus, not the child-specific corpus, whereas the user's age is considered a critical factor in creating a dictionary. Different characteristic requires a piece of different information from a dictionary. Children and adults need different information from a dictionary (Tono, 2012). Therefore, children need different dictionaries from adults.

Another advantage possessed by Google Translate compared to Kamusku is the information on examples of the use of words in sentences. Examples of using words in this sentence are essential information that a dictionary must-have because it can help students learn a word according to its context. The results of research support it by Farina, Vrbinc, & Vrbinc (2019)their look-up abilities, and their perceptions of the utility and quality of definitions and illustrative examples. Students were given nine contexts containing a clearly-marked common word used in an infrequent sense; they had to locate the relevant sense in the online Merriam–Webster Learner's Dictionary (MWLD which state that students need complete sentences rather than phrases. The examples of how to use words in sentences help students to understand the vocabulary. However, the vocabulary contained in the example sentences on Google Translate contains a lot of vocabulary at the K3 level, while learning English at the junior high school level only learns vocabulary at the K1 level. So that this information does not help students much in understanding the vocabulary they are learning.

From the results of the discussion above, it can be concluded that the two dictionaries have their respective advantages and disadvantages. In general, Google Translate is superior to Kamusku. However, Google Translate still does not meet the user's needs, and it is inappropriate to be used in learning English at Junior High School Level. Students feel overwhelmed by too much information listed in Google Classroom. This inappropriacy affects the effectiveness of the dictionary because this condition makes students confuse when using the dictionary.

5 References

Alhaisoni, E. (2016). EFL Teachers' and Students' Perceptions of Dictionary Use and Preferences *International Journal of Linguistics*, *8*(6), 31. https://doi.org/10.5296/ijl.v8i6.10267

Alharbi, M. A. (2016). Using different types of dictionaries for improving EFL reading comprehension and vocabulary learning. *JALT CALL Journal*, *12*(2), 123–149.

Asgari, A., & Mustapha, G. Bin. (2011). The type of vocabulary learning strategies used by ESL students in University Putra Malaysia. *English Language Teaching*, *4*(2), 84.

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. New York: Oxford University Press.

Barham, K. A. (2017). The Use of Electronic Dictionary in the Language Classroom: The Views of Language Learners. *Online Submission*.

Chaer, A. (2007). *Leksikologi dan Leksikografi*. Jakarta: Rineka Cipta.

Chan, A. Y. W. (2017). The effectiveness of using a bilingualized dictionary for determining noun countability and article selection. *Lexikos*, *27*(1), 183–213.

Cimermanová, I. (2012). Corpus vs dictionary in EFL classes. *English Matters III*, 65–73.

Ebanéga, G. M. E., & Moussavou, F. T. (2008). A Survey of the Dictionary Use of Gabonese Students at Two South African Universities. *Lexikos*, *18*(1).

Farina, D. M. T. C., Vrbinc, M., & Vrbinc, A. (2019). Problems in Online Dictionary Use for Advanced Slovenian Learners of English. *International Journal of Lexicography*, *32*(4), 458–479. https://doi.org/10.1093/IJL/ECZ017

Golavar, E., Beikian, A., Nooramin, A. S., & Firoozkoohi, S. (2012). Monolingual vs. bilingual dictionaries for learning technical terms. *Journal of Basic and Applied Scientific Research*, *2*(5), 452–457.

Granger, S. (2012). Electronic lexicography: From challenge to opportunity (pp. 1–11). https://doi.org/10.1093/acprof:oso/9780199654864.003.0001

Heuberger, R. (2016). Corpora as game changers: The growing impact of corpus tools for dictionary makers and users: Corpus tools have created a paradigm shift in English lexicography. *English Today*, *32*(2), 24–30.

Jackson, H. (2013). *Lexicography: an introduction*. Routledge.

Kwary, Deny A. (2018). A corpus and a concordancer of academic journal articles. *Data in Brief*, *16*, 94–100.

Kwary, Deny Arnos. (2003). Creating a Technical Vocabulary Wordlist and Web-based Materials: The First Step for Teaching and Learning ESP.

Nesi, H. (2013). Researching users and uses of dictionaries. *The Bloomsbury Companion to Lexicography*, 62–74.

Nesi, H., & Bae, S. (2014). Korean and English 'dictionary'questions: what do the public want to know?

Nkomo, D., & Madiba, M. (2012). The Compilation of Multilingual Concept Literacy Glossaries at the University of Cape Town: A Lexicographical Function Theoretical Approach. *Lexikos; Vol 21 (2011) DO - 10.5788/21-1-41*. Retrieved from http://lexikos.journals.ac.za/pub/article/view/41/49

Nugraha, D. N. S. (2020). Grammatical and Diction Inaccuracy in English—Indonesian Translation on Google Translate. In *International Conference on Educational Psychology and Pedagogy-" Diversity in Education"(ICEPP 2019)* (pp. 35–39). Atlantis Press.

Oppentocht, L., & Schutz, R. (2003). Developments in electronic dictionary design. In P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 215–227). Amsterdam: John Benjamins.

Rezaei, M., & Davoudi, M. (2016). The Influence of Electronic Dictionaries on Vocabulary Knowledge Extension. *Journal of Education and Learning*, *5*(3), 139–148.

Şevik, M. (2014). University Prep-school EFL Learners' Dictionary Ownership and Preferences. *14th Language, Literature and Stylistics Symposium*, *158*(December 2014), 226–232. https://doi.org/10.1016/j.sbspro.2014.12.080

Tan, K. H., & Woods, P. C. (2008). Media-Related Or Generic-Related Features In Electronic Dictionaries: Learners? Perception And Preferences. *GEMA Online Journal of Language Studies*, *8*(2), 1–17. Retrieved from http://www.fpbahasa.ukm.my/linguistics/Gema/page1_17.pdf

Tono, Y. (2012). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension* (Vol. 106). Walter de Gruyter.

Wenner, L., & Sköldberg, E. (2019). Folkmun.se: A Study of a User-Generated Dictionary of Swedish. *International Journal of Lexicography*, 1–19. https://doi.org/10.1093/ijl/ecz019

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The Oxford Children's Corpus: Using a Children's Corpus in Lexicography1. *International Journal of Lexicography*, *26*(2), 190–218. https://doi.org/10.1093/ijl/ecs017

Zheng, H., & Wang, X. (2016). The use of electronic dictionaries in EFL classroom. *Studies in English Language Teaching*, *4*(1), 144–156.

THE ANALYSIS OF STRUCTURE OF THREE INDONESIAN MONOLINGUAL DICTIONARIES

Azhari Dasman Darnis, Umi Kulsum
National Agency for Language Development and Cultivation, Indonesia;
UIN Syarif Hidayatullah, Jakarta
azhari.dasman@kemdikbud.go.id; umikulsumfah@uinjkt.ac.id

Abstract

Kamus Umum Bahasa Indonesia (KUBI) or Indonesian General Dictionary, *Kamus Bahasa Indonesia* (KBI) or Indonesian Dictionary, and *Kamus Besar Bahasa Indonesia* (KBBI) or Indonesian Comprehensive Dictionary are Indonesian monolingual dictionaries compiled by the National Agency for Language Development and Cultivation, the Indonesian government body in charge of developing and cultivating Indonesian and regional languages and literatures. The first two dictionaries did not live long. KUBI stopped publishing after its third edition in 2003. KBI as the second dictionary compiled by the institution marked its inaugural launch as the termination of the publication. The circulation of KBI is limited and was out of print after the first publication in 1983. KBBI, the last dictionary compiled by the same institution, which development has continued until now is updated regularly, every six months. KBBI is published in print, online, and offline mobile versions. This study is addressed to investigate the structural aspects of three dictionaries to come with the results in the profile of their formats and component parts. To come with the results the investigation will be applied on front matter and back matter of dictionaries as well as on the macro-and microstructure known as middle structure. This study is a metalexicography study that makes use of the qualitative descriptive method. To bring about the results purposed sampling of entries from each dictionary was exercised. The research results that of the three dictionaries the KBBI are the most comprehensive dictionary, while the KBI is the readiest dictionary to answer the need for a standard dictionary. As for KUBI, it is the most synchronic.

Keywords: KBBI, KUBI, KBI, dictionary structure, microstructure

1. Introduction

There are three dictionaries of Indonesian language that are compiled and managed by the government agency currently named Badan Pengembangan dan Pembinaan Bahasa or Badan Bahasa, an Indonesian government agency in charge of developing and cultivating regional languages and literatures of Indonesia (Indonesian Law No. 24 Anno 2009) or known as Badan Bahasa. Of the three dictionaries, only one which management and development continue up until now, namely Kamus Besar Bahasa Indonesia (KBBI) or The Comprehensive Dictionary of Indonesian Language. Two others, Kamus Umum Bahasa Indonesia (KUBI) or The General Dictionary of Indonesian Language (KUBI) compiled by W.J.S. Poerwadarminta and Kamus Bahasa Indonesia (KBI) or Indonesian Dictionary of Sri Timur Suratman and the team, not longer printed and published. KUBI was not updated again after its third edition was published in 1976, but the dictionary was still printed until 2003 for a limited user. Though the KUBI final edition has been adapted the Indonesian latest spelling rules (EYD) and has added hundreds of new entries (Alwi, 1976: viii).

The Indonesian Dictionary (KBI) is the second Indonesian monolingual dictionary compiled by the same institution. Most of the teams working on KUBI and KBI are the same. The preparation of the dictionary had begun two years before the last edition of KUBI was published in 1974. KBI's inaugural launch in

1983 was marked as the termination of its publication. The dictionary is only circulating among limited people.

KBBI is the last Indonesian monolingual dictionary compiled by the institution, which development continues until now. Since its last edition (2016) the dictionary has been published on three platforms: print, online, and offline mobile version. The dictionary is also available in print and digital of Braille editions. Besides updated regularly, every six months, the dictionary also has a digital database. The database is regularly updated as the updating of its content is applied. However, the massive penetration of KBBI among users has not diminished the users' hopes of the existence of a standard dictionary which is the main reason behind the termination and growing Indonesian language dictionaries in the institution.

The public's need for a dictionary other than KBBI has not been extinguished. People's longing for a standard dictionary that only contains standard entries has been echoed since the Third Indonesian Language Congress in 1978. One of the points of general conclusion in the Section of the Development of The Indonesian Language with Relation to the Field of Linguistics mentioned that "The standard Dictionary of Indonesian language needs to be published and disseminated immediately. For the purpose research in the field of lexicology needs to be carried out and experts in various fields of science are should be involved" (Agency for Language Development and Cultivation, Indonesian Congress Decision Group I-IX, 2011: 33).

Following the issue, this study will examine the three dictionaries above from a structural perspective (Atkins and Rundel 2008, Hartmann 2001, and Jackson 2003). This study looked at all three dictionary structures, ranging from front to end matter. The middle matter section containing macro-and microstructure is the part that got a considerable portion of study because it is the most important part of a dictionary (Hartmann, 2001:178). Through microstructures study profile formats and components parts of the dictionary structure will be drawn (Hartmann, 2001: 57). Into the results added the information about style guides and template entries in particular lexical sets as well.

2. Method

2.1. Methodology

The study takes a larger portion on the microstructure section conducted by, first, choosing an example of the entries from all three dictionaries. The example was chosen based on the division of word classes in the dictionaries. Each single word class is represented by one randomly drawn entry from every single dictionary. Since the new word class only available in KBI and KBBI, so the division of word classes in both dictionaries is used, they are namely: nouns, verbs, adjectives, adverbs, numerals, particles (KBI: xxxvi), and pronouns (KBBI V, 2016: xxxix).

As the sample, seven entries representing formal functions are selected , namely *Senin* (n), *berontak* (v), *curiga* (adj), *seluruh* (num), *barangkali* (adv), *ketika* (part), and *aku* (pron). Those entries are analyzed to find out thirteen microstructure components proposed by Atkins and Rundel (2008) Hartmann (2001) and Jackson (2003), as follow *syllable, pronunciation, derivation, word class, senses, definition, example, usage, run-ons, idiom, phrasal verb, etymology,* and *collocation example,*

Table 1. Entry Samples

1.	**noun**	*Senin, Senen*	Monday	5.	**adverb**	*barangkali*	perhaps
2.	**verb**	*berontak*	to fight	6.	**particle**	*ketika*	when
3.	**adjective**	*curiga*	suspicious	7	**pronoun**	*aku*	I/me
4.	**numeral**	*seluruh*	whole				

These entries are analyzed to find out the type of information or components, style guides, and the entry templates included. Thirteen microstructure components in the structure would be analyzed (Atkins and Rundel 2008, Hartmann 2001, and Jackson 2003). These components are as follows.

Table 2. Microstructure components

sy	=	Syllable	us	=	Usage
pro	=	Pronunciations	ro	=	run-ons
der	=	Derivation	id	=	Idioms
wc	=	word class	pv	=	phrasal verbs (if they are not included as headwords)
se	=	Senses	ety	=	Etymology
def	=	Definitions	ce	=	collocational example
ex	=	Examples			

To find out whether the dictionary makes use of a particular entry template or not is carried out by checking a group of entries that belong to a particular lexical set which is done by examining the entries of the days of the week in all three dictionaries.

3. Result

Kamus Ekabahasa Bahasa Indonesia Badan Bahasa

Badan Bahasa has launched several dictionaries. Those dictionaries are compiled for various purposes that generally can be divided into two types. First, general-purpose dictionary. Second, a special purpose dictionary. All three dictionaries above belong to the first type. The special purposes dictionaries such as the Indonesian dictionary for students, dictionaries of sciences, etymology dictionaries, etc. These specific dictionaries are excluded from the discussion of this research.

a. *Kamus Umum Bahasa Indonesia* (KUBI)

KUBI is designed to be a practical Indonesian language dictionary to meet the need in understanding all kinds of reading material (KUBI, 1952: 5) or to eliminate obstacles during reading (text-related problems). The great number of reading materials and the absence of reference works at that time underlay Poerwadarminta to compile KUBI. In addition, at that time, he was compiling an Indonesian-Dutch dictionary, along with A. Teew, so the compiling of both dictionaries is conducted side by side based on the same data.

Both dictionaries, Indonesia-Dutch Dictionary and KUBI are compiled based on the same material and processed in the same way and method. "Along, because both are based on the same ingredients and processed in the same scene. Exchanging roads, because each goes on its own way, it is in line with the circumstances and environment of the user" (KUBI, 1953: 5). So, KUBI is the twin of the Indonesian-Dutch Dictionary that has been published earlier. All the disadvantages and advantages delivered in the foreword of the latter applied to the first as well (KUBI, 1953: 5).

KUBI is based on, according to Poerwadarminta, two types of data, an old and a new one. The old data is collected from the treasury of Malay old cultural and literary treasures such as tales, *pantuns*, poems, and so on that are found in reading books and lessons at that time (KUBI, 1953: 5). KUBI heavily bases its data on literary works such as Layar Terkembang novelette and all publications that existed in the 1920s or before. Meanwhile, the new data extracted from words and terms of Western languages (Europe), Arabic, Latin, Chinese, Sanskrit, and regional languages (Sundanese, Minangkabau, and Palembangese).

The entries of KUBI exclusively have a high frequency of use and distribution. It is measured by the use of each word in at least five places, namely Medan, Batavia, Surabaya, Ambon, and Makassar which represent major cities in four major islands in Indonesia and five different regions (Lasmiah, 1980). The frequency use of such collected lexicography data was also tested through existing publications. Each word has been used at a minimum, in five publications, either magazines or books, and by five different authors. Poerwadarminta spent seven years collecting data and another three years compiling it in the macro-and microstructures.

Poerwadarminta's Dictionary is designed to be a descriptive, practical, and simple one (Poerwadarminta, 1953) and based on the available corpus. As for definition, for instance, it has been based on the meaning of the word in its context. Likewise, examples are taken from excerpts in books, magazines, and newspapers. The compilation of KUBI can be justified scientifically because of the adaption of lexicography rules (KBBI, 2008: xxxviii).

b. *Kamus Bahasa Indonesia* (KBI)

The Indonesian Dictionary or KBI began to be compiled in 1974 and published nine years later in three volumes. The first volume containing the entry A--J, the second volume containing the entry K--S, and the third volume containing the rest. The publication of the dictionary was carried out to accommodate the development of the Indonesian language and Indonesian more systemic vocabulary and adopt the development of advanced lexicography rules (KBI, 1983: vii). Although only circulating in a limited number of users, KBI was expected to be a comprehensive or a standard dictionary in the future. However, considered not meet that expectation, Badan Bahasa then formed a brand new team to compile another advanced Indonesian monolingual dictionary.

c. *Kamus Besar Bahasa Indonesia* (KBBI)

Kamus Besar Bahasa Indonesia (KBBI) first edition published in 1988, it has 62,000 entries. The number of its entries then increased by about 10,000 over three years in the second edition (1991). Its third edition, published in 2001, contains 78,000 entries and the fourth edition published seven years later has more than 92,000 entries. The fifth edition was released for the first time in 2016. The latest launched in three formats: printed, online, and offline mobile versions. It is regularly updated twice a year. As of April 2019, it has more than 110,000 entries. Moeljadi et al. (2017) describe the creation of the database as well as the database structure. Kamajaya et al. (2017) explain the online KBBI in detail. (Moeljadi, et.al in Asialex 2019 Prosiding:174).

4. Analysis and Discussion

4.1 Front and Back Matter

Kamus Umum Bahasa Indonesia (KUBI)

KUBI's front matter consists of the preface and the instructions of use. The instruction uses the phrase "some hints" because they contain several things. The first thing is related to the alphabet and spelling used. A brand new spelling rule called Spelling of the Republic of Indonesia was just introduced replacing Van Ophuijsen Spelling. There are many books, readings, and publications that still make use of the latter spellings so it is necessary to disseminate the use of new spelling rules via KUBI.

The second instruction is related to the variations in pronunciation and spelling in Indonesian due to new spelling rules. Poerwadarminta divides these variations into groups of vowels and consonants. The first variation are, for example, "*tentara* and *tentera*, *anggauta* and *anggota*", while consonant variations such as "*hayal* and *khayal* or *zohor* and *lohor*". The instruction also contains a glimpse of the spelling of Malaysia Malay Language under the title "Edjaan Semenandjung Tanah Melayu" for practical purposes of Malay language users in Malaysia.

Such brief instructions were necessary to bridge the spelling differences between Indonesian and Malaysian Malay. Whether KUBI encourages what so-called a joint spelling (*ejaan bersama*) of Indonesia and Malaysia is not known yet. Certainly, the joint spelling of the two countries was compiled not long after the publication of the third or seventh of KUBI's first edition. However, the joint spelling failed to be legalized due to the political issues between Indonesia and Malaysia at the time.

The instructions also provide a brief practical explanation of morphology of words derived from

Jakarta-Malay and Javanese. This is interesting because the entries of KUBI are not limited derived from those languages but also from the Minangkabau language, Sundanese, Palembangese. In this regard, at least two questions arise. Firstly, is it because the morphology of both languages considered extraneous, whereas the morphology of other regional languages in Indonesia, such as Minangkabau, is not? Secondly, does it because of the absorption Jakarta-Malay and Javanese words into Indonesian easier than that of others?

The final instructions of use explains the macro-and microstructure of the dictionary. This section explains that the label of the origin of words, such as A (Arabic), and Djw (*Djawa*), are used to state that those words are not considered to be common Indonesian. The argument is not entirely true because some entries labeled A, for example, have been entered into the Malay language a long time ago and inscribed in ancient manuscripts. Such words can be considered to have entered into Bahasa Indonesia both in terms of form and meaning, such as "mizan" and "mistar". Supposedly, other words are still considered as the outsider because their forms not matche the spelling of the Indonesian language as "*mintaku'lburudj*" (KUBI, 1953:460). The word is morphologically Arabic.

The labels for dictionary user easiness consist of --, as a substitute for a word, ~ substitutes derivational words, - cross-references, = similar with, and crosses (†) to mark words that are still disputed (misheard, misquoted, misreading, etc.), rarely found, found only in restricted area (regional languages, etc.), obsolete (dead), etc. (Ibid: 10)

Instructions on labels or abbreviations consist of instructions on the etymology of the language, namely Arabic, Jakarta Malay, Javanese, Europe, Latin, Minangkabau, Palembangese, Sundanese, Sanskrit, and Tionghwa. (Ibid.)

The compilation of KUBI already rests on the written corpus in the form of citations (see Atkins and Rundell, 2008: 61) either for a definition or for examples.

"In general, the examples are quoted from the reading materials used as the basis for compiling this dictionary". (KUBI, 1952; 7)

"... derived from the regional language and usually used in Indonesian books, ..." (KUBI, 1952; 7)

The back matter as Atkins and Rundell said (2008:177) often includes lists of verb tables, numbers, weights and measures, chemical elements, Roman numerals, etc, but it may also provide maps, diagrams, and other material geared to the needs of the target user not found yet in KUBI.

KUBI's end matter sections are alphabetically arranged and not all filled with abbreviations as stated in the title, but contain entries as well. The example of abbreviations are C.H.T.H (*Chung Hua Tsung Hui*), B.T.I (*Barisan Tani Indonesia*), and S.G.A (*Sekolah Guru Atas*). It contains also some clipping such as K., Kg. (*Kandjeng*), title as R.Ngt. (*Raden Nganten*), and wk. (*wakil*) and abbreviations that often found in writing at that time such as "j.a.d. (*jang akan datang*)", "upm. (*umpama*.) "sjd. (*sampai dengan*)", and so on. The entries that included in end matter section are one, two, three, or four-letter abbreviation that contains short definition or glosses, such as:

(Perantjis: à) *tiap-tiap*; se ……; *6 meter kain à Rp. 3.~;* 2 are (100m^2)

c.i.f : (Inggeris) cost insurance freight (*ongkos, asuransi dan pengangkutan sudah terhitung dalam harga*)

p., pag. : (Lat) pagina (*muka halaman*)

ult. : ultimo (*pd achir bulan*)

Most of such abbreviations such is no longer found in the next two dictionaries. Those abbreviations were very typical of the 1940s and were strongly associated with the social and political situation at the time.

The first and second examples of entries above are not included in the middle matter, but the definition of the third “pagina” did. However, the lemma “p.” or “ pag” was excluded. The last entry “ultimo” is excluded as well from the middle matter but included in the third edition published in 1976. After the dictionary remanaged by by Badan Bahasa the end matter under the title “Abbreviation” no longer exists until the last edition of the 2003 print.

Not all the entries and abbreviations available in the end matter attached to the middle matter. The entries “*mangkunegaran, raden mas, raden panji,* and *jang mulia*”, for example, are excluded in the next edition middle matter. However, the entry “s.w.t: *subhanahu wa ta’ala*”, for example, is available under the headword “Subhana A”, but “Subhana” alone not a well-known word in Indonesian.

KUBI did not list books or publications used as references. The list of references never showed up, even after the management of the dictionary already handed over Badan Bahasa.

Kamus Bahasa Indonesia (KBI)

The need to accommodate the development of the systematic Indonesian vocabulary is illustrated from the front matter. The section explained the new spelling rules of the Indonesian language and clarified the morphological rules, the rules of absorption of foreign language vocabulary into Indonesian, the standard and non-standard words, as well as the syllable system which are used in dictionaries.

All the conventions used in the dictionary are explained in the instruction of use section (Atkins and Rundell, 2008: 177). Included in this section is the use of labels, MWE, the use of chemical formulas, abbreviation entries, and so on. At the end inserted a list of references used to compile the KBI.

The KBI does not have any information on the end matter, but on the front matter, it has a reference section containing a list of books and publications used as a database for compiling the dictionary. It is located on the last section of the front matter. The front matter of KBI is enriched by new spelling rules, morphological rules, the rules of absorption foreign language, and non-standard words adopted from regional and foreign languages.

Another section on the front matter is about the new role of KBI. It has another function as the instrument for the development of the Indonesian language. It highlights its prescriptive purposes because Badan Bahasa, is responsible for fostering, developing, and cultivating a practical, rich, and highly expressive Indonesian language to become an authoritative language of education and unity as the mandate of the Indonesian Language Congress the Third (The Decision of Indonesian Language Congress,1978) to uphold the position and function of the Indonesian language in the country (KBI, 1983: xvi).

Kamus Besar Bahasa Indonesia (KBBI)

Kamus Besar Bahasa Indonesia or KBBI is the most complete in structure amongst the three dictionaries. The dictionary equipped with front, back, and middle matter. KBBI first published was in 1988 and has published its fifth edition until now. The dictionary has also been published on a variety of platforms: print, online, and offline mobile versions. Besides, KBBI is also published in print and electronic Braille editions.

The front matter section typically contains things as foreword and acknowledgments, introduction to the dictionary, abbreviations, labels, and code (Atkins and Rundell, 2008: 176). In addition, the front matter section of KBBI offers mini-essays on the history of lexicography in Indonesia. In the end matter, KBBI has important information according to the target user in line with Atkins and Rundell (Ibid) statement that end matter may also provide material geared to the needs of the target user. By this, KBBI provides regional words and expressions, foreign words and expressions, regional characters in Indonesia, abbreviations and acronyms, national and international holidays, stars and honor marks, country names, capitals, languages, and currencies, provincial and district names as well as broad descriptions of regions and their inhabitants, signs, and symbols, sizes and scales, and so on. (KBBI, 2016:1583—1701)

4.2 Microstructure of KUBI, KBI, and KBBI

Discussions about the component of information in the microstructure, style guide, and template entry of the three dictionaries are presented below.

4.2.1 Component of Information in Microstructure

Overall the amount of information contained in the microstructures in all three dictionaries varies. KBBI is the highest of the three and KBI is the second. The amount of information on KBI microstructures is more than that of KUBI.

Table 3, Microstructure component

Entri	**KUBI**	**KBI**	**KBBI**
Senin, Senen	2	5	6
berontak	4	6	7
curiga	4	6	6
seluruh	4	6	7
barangkali	3	5	5
ketika	4	7	8
aku	3	5	5
Average	3,4	5,7	6,2

Table 4, Microstructure component on KUBI

Entries	**Senin, Senen**	**berontak**	**curiga**	**seluruh**	**barangkali**	**ketika**	**aku**	
Kind of information								
syllable	x	x	x	x	x	x	x	
pronunciation	x	x	x	x	x	x	x	
derivation	-	v	v	v	-	v	v	
word class	x	x	x	x	x	x	x	
Senses	v	v	v	v	v	v	v	
Definitions	v	v	v	v	v	v	v	
Examples	x	v	v	v	v	v	x	
Usage	x	x	x	x	x	x	x	
Idioms	x	-	-	-	-	x	-	
phrasal verbs (if they are not included as headwords)	x	-	-	-	-	x	-	
etymology	x	x	x	x	x	x	x	
collocational example	x	x	x	-	x	x	x	
Amout	2	4	4	4	3	4	3	24

Table 5, Microstructure component on KBI

Entries	**Senin**	**berontak**	**curiga**	**seluruh**	**barangkali**	**ketika**	**aku**	
Kind of Informations								
syllable	v	v	v	v	v	v	v	
pronunciation	x	x	x	x	x	x	x	
derivation	-	v	v	v	-	v	v	
word class	v	v	v	v	v	v	v	

Senses	x	v	v	v	v	v	v	
Definitions	v	v	v	v	v	v	v	
Examples	v	v	v	v	v	v	x	
Usage	x	x	x	x	x	x	x	
Idioms	v	-	-	-	-	-	-	
phrasal verbs (if they are not included as headwords)	-	x	x	-	-	v	-	
etymology	x	x	x	x	x	x	x	
collocational example	x	x	x	x	x	x	x	
Amount	5	6	6	6	5	7	5	40

Table 6, Microstructure component on KBBI

Entries	**Senin**	**berontak**	**curiga**	**seluruh**	**barangkali**	**ketika**	**aku**	
Kind of Informations								
syllable	v	v	v	v	v	v	v	
pronunciation	v	v	x	v	x	v	x	
derivation	-	v	v	v	-	v	v	
word class	v	v	v	v	v	v	v	
Senses	v	v	v	v	v	v	v	
Definitions	v	v	v	v	v	v	v	
Examples	x	v	v	v	v	v	x	
Usage	x	x	x	x	x	x	x	
Idioms	v	-	-	-	-	v	-	
phrasal verbs (if they are not included as headwords)	-	-	-	-	-	-	-	
etymology	x	x	x	x	x	x	x	
collocational example	x	x	x	x	x	x	x	
Amount	6	7	6	7	5	8	5	44

4.2.2. Template Entry

How the three dictionaries organize their entries into the lexical set of days of the week as follows.

Table 7, Lexical Set Template Entry of KUBI

Headwords	**Definitions**
Senin	hari yg kedua -> Senen, Isnain
Selasa	Selasa, **hari** ~ hari yg ketiga
Rabu	(**hari** ~): hari yg keempat; -> Arba’a
Kamis	**hari** ~; hari yg kelima -> Kemis
Jumat	hari yg keenam -> djuma’at
Sabtu	**Sabtu, hari** ~: hari yg ketujuh
Ahad	**ahad** A: 1 …; 2 (hari ~), hari Minggu; 3 minggu
Minggu	**minggu:** 1 *(hari ~),* hari Ahad

As presented in the table, KUBI has an entry template entry for days of the week, but *Senin* (Sunday). The definition of *Senin* is synonymous and circular.

Table 8, Lexical Set Template Entry of KBI

Headwords	**Definitions**
Se.nin	*n* hari yg kedua (sesudah Minggu)
Se.la.sa	**Selasa**, (**hari --)** *n* hari yg ketiga dl satu minggu pd penanggalan Masehi; hari sesudah Senin
Ra.bu	*n* hari keempat dl seminggu
Ka.mis	hari yg kelima
Ju.mat	*n* **1** nama hari yg keenam
Sab.tu	*n* nama hari ketujuh
a.had	*Ar n* **1** satu; esa; **2** (hari) Minggu
ming.gu	*n* **1** hari ke-1; Ahad

As we can see, there is no consistency in the definitions of the days of the week go in KBBI. None of the definition models could be taken as a template entry because each entry has its own way of definition. Note, the entry "Senin" is located between entry **sen.duk** and **se.ne.wen**, it should be between **se.ni.man** and **se.ni.or.**

Table 9, Lexical Set Template Entry of KBBI

Headwords	**Definitions**
Se.nin	*n* hari *ke-2* dalam jangka waktu satu minggu; *Senen
Se.la.sa	*n hari ke-3 dalam jangka waktu satu minggu*
Ra.bu	*n hari ke-4 dalam jangka waktu satu minggu; *Rebo*
Ka.mis	*n hari ke-5 dalam jangka waktu satu minggu*
Ju.mat	*n hari ke-6 dalam jangka waktu satu minggu*
Sab.tu	*n hari ke-7 dalam jangka waktu satu minggu*
Ahad	*n hari pertama dalam jangka waktu satu minggu; Minggu; *Akad*
ming.gu	*n (ditulis dengan huruf besar) hari pertama dalam jangka waktu satu minggu; Ahad*

KBBI has two entry templates model for the lexical set of the days of the week. The first entry template for *Senin--Sabtu*. Second, a template entry for *Ahad* and *minggu*. In particular, for the last two entries, the first alphabet of the lemma starts in lowercase, not as it should be. The information of the usage of the capital letter is described in a gloss in parenthesis because one of the two names for the same day is a nonstandard variant, initial use of lowercase letters is possible for such purposes.

4.2.3 The Accommodation of a Non-standard Entry

Besides the inconsistency of the use of entry templates, the tables above also inform the accommodation of non-standard entries in every dictionary. Of the three tables representing the three dictionaries, only KUBI and KBBI accommodate non-standard entries. KUBI lists five non-standard entries of the lexical set containing eight entries. Meanwhile, KBBI lists three non-standard entries of a lexical set containing eight entries. Only KBI excludes nonstandard entries, in line with the instructions presented in the front matter section of the dictionary.

Nonstandart Entry of KUBI

Senin hari yg kedua -> Senen, Isnain

Rabu (**hari** ~): hari yg keempat; -> Arba'a

Kamis **hari** ~; hari yg kelima -> Kemis

Jumat hari yg keenam -> djuma'at

Nonstandart Entry of KBBI

Se.nin *n* hari *ke-2* dalam jangka waktu satu minggu; Senen

Ra.bu *n hari ke-4 dalam jangka waktu satu minggu;* Rebo

Ahad *n hari pertama dalam jangka waktu satu minggu; Minggu;* Akad

5. Conclusion

This research successfully concluded that among the three dictionaries, KUBI, KBI, and KBBI, the latter has the most information on the structure. In the middle-matter section KBBI includes the component information of formal comments consisting of pronunciation, syllables, and word classes. The information of lexical comment includes definitions, meanings, examples of usage, derivations, MWE, idioms, expressions, and proverbs. KBBI has also maximized the benefits of the front and post matter. KBI potentially accommodates the need for standard dictionaries, while KUBI can be proposed as a synchronous dictionary.

References

Adiwimarta, Sri Sukesi (1983). *Kamus Bahasa Indonesia.* Jakarta: Pusat Pembinaan dan Pengembangan Bahasa

Alwi, Hasan (1976). *Kamus Bahasa Indonesia.* Jakarta: Pusat Pembinaan dan Pengembangan Bahasa. Edisi K-2, Cet. Pertama

Atkins, B.T Sue and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Badan Pengembangan dan Pembinaan Bahasa (2011). Kumpulan Putusan Kongres Bahasa Indonesia I—IX, 2011. Jakarta: Badan Pengembangan dan Pembinaan Bahasa

Hartmann, R.R.K. (2001). *Teaching and Researching Lexicography*. Essex: Pearson Educational Limited

Hartmann, R.R.K and James, G. (2001). *Dictionary of Lexicography.* Routledge

Jackson, H. (2002). Lexicography an Introduction. Routledge

Kementerian Pendidikan dan Kebudayaan. (2016). *Kamus Besar Bahasa Indonesia.* Balai Pustaka: Jakarta

Lasmiah, Putu. 1980. Tokoh Nasional W.J.S Poerwadarminta. Jakarta: Departemen Pendidikan dan Kebudayaan, Pusat Penelitian Sejarah dan Budaya, Proyek Inventarisasi & Dokumentasi Sejarah Nasional,

Poerwadarminta, W.J.S. (1953). *Kamus Umum Bahasa Indonesia.* Jakarta: Balai Pustaka

http://badanbahasa.kemdikbud.go.id/lamanbahasa/artikel/3438/sejarah-kamus-besar-bahasa-indonesia accessed on 24th of May 2021

http://badanbahasa.kemdikbud.go.id/lamanbahasa/artikel/3418 accessed on 24th of May 2021

USER-PROFILE DEFINITION AND INTERFACE DESIGN: THE CASE OF THE *PARALYMPIC DICTIONARY*

Bruna da Silva[1;2], Ana Luiza Vianna[2], Sandra de Oliveira[2], Rove Chishman[2], Gilles-Maurice de Schryver[1]
Ghent University, Belgium[1]; UNISINOS University, Brazil[2]
bruna.dasilva@ugent.be; alvianna@edu.unisinos.br; sandraoliv@edu.unisinos.br; rove@unisinos.br; gillesmaurice.deschryver@ugent.be

Abstract

Compiling a dictionary is more than developing contents; it requires, first and foremost, planning. Among the decisions that must be made during the dictionary-planning or pre- lexicographical stage (Atkins & Rundell 2008: 18-44), one of the most important is the user- profile definition, which is the key for creating a lexicographical resource that achieves the users' expectations and is thus functional. Given the relevance of this matter, the present work aims to design a user profile for the *Paralympic Dictionary* (under development), and to investigate its implications for the dictionary-making stage. This dictionary is part of a broader set of lexical resources whose goal is to describe sports' lexicon based on Frame Semantics (Fillmore 1982, 1985). Although the two previous dictionary outputs from the SemanTec research group — viz. *Field*, a football expressions dictionary (Chishman 2014) and *Dicionário Olímpico*, a dictionary of Summer Olympic sports (Chishman 2016) — had a target user group in mind (cf. Chishman *et al.* (2014), dos Santos & Chishman (2015), Chishman *et al.* (2018), Chishman *et al.* (2019)), one might say it was a rather basic one whose implications were not widely discussed nor well incorporated into the design of the dictionaries. Focusing on *Paralympic Dictionary*'s envisaged users, the current work deepens and expands on the earlier discussion about target-audience definitions in the context of the SemanTec research group. In contrast with the other two resources, the user-profile definition for the *Paralympic Dictionary* intends to go beyond a broad notion of 'lay audience', determining the specific user groups and their specific needs. Regarding the compilation of the *Paralympic Dictionary*, the user-profile definition will help determine the functions, content and structure of the dictionary and enable the evaluation of which aspects of the previous dictionaries one should maintain or redesign, and to reflect on the type of inclusion of elements as well as on the adoption of digital access policies.

Keywords: user profile, target audience, pre-lexicographical stage, frame-based dictionary, Paralympic sports.

1 Introduction

The definition of the intended target audience serves as a basis for the planning and compilation activities of any type of dictionary as it guides and directs the lexicographer in establishing the functions, defines the contents, and designs the dictionary structure (cf. Gouws 2011, 2018, 2020). In the development of dictionaries based on frames, the definition of the user profile is as important as the theoretical notions from Frame Semantics, and both exert a shared influence on the planning of the dictionary, a phenomenon that can be observed in the dictionaries developed by the SemanTec (i.e., Semantics and Technology) research group.

In general, the research developed by SemanTec seeks to investigate and explore the interaction between Frame Semantics (Fillmore 1977, 1982, 1985) and Internet Lexicography (or Online Lexicography (Gouws, 2018)) through the description of the lexicon of special fields (cf. Gouws, 2020). To date, this research has culminated in the publication of the dictionaries *Field – Dictionary of football expressions*[1] (Chishman, 2014) and *Olympic Dictionary*[2] (Chishman, 2016). Although there are differences between them, both tools adopt the notion of frames to describe the lexicon of special fields (soccer and Summer Olympic sports, respectively).

In addition to the organization format, the dictionaries also share similarities regarding target audience definitions: during the resources' planning process, the SemanTec team adopted a relatively broad and wide user profile (cf. Chishman *et al.* (2014), dos Santos & Chishman (2015), Chishman *et al.* (2018), Chishman *et al.* (2019)) from which lexicographic decisions were made.

Considering the centrality of the user-profile definition for the planning and development processes of dictionaries, the present work aims to problematize the implications that the adopted user concept brought to the design of the interfaces for the *Field* and *Olympic* dictionaries. Additionally, this work also seeks to complement the documentation of the steps followed by SemanTec during the compilation of these tools. This discussion arises from the revision of the *Olympic Dictionary* and constitutes, above all, an exercise in theoretical and methodological reflection (and why not self-criticism?) intended to identify and correct flaws that, to some extent, originated from the established user profile. In addition to the contribution to the reformulation of the dictionaries already published, the consequences of such reflection extend to future works and will be valuable for discussions related to the *Paralympic Dictionary*, the current project at SemanTec.

In order to situate the reader in relation to the various projects, Section 2 presents the existing *Field* and *Olympic* dictionaries and their respective user profiles based on publications by the SemanTec group. Section 3 problematizes the user profiles of the *Field* and *Olympic* dictionaries in order to serve as a basis for the description of the *Paralympic Dictionary*'s user profile. Section 4, finally, presents a brief conclusion.

2 The *Field* and *Olympic* dictionaries: Target user groups and their broader implications

The *Field* and *Olympic* dictionaries are endeavours in the interface between specialised lexicography and Frame Semantics: *Field* is a tridirectional trilingual dictionary (Portuguese, English and Spanish) that describes the football lexicon; the *Olympic Dictionary* is a unidirectional bilingual resource (Portuguese with translation equivalents and examples in English for the lexical units) that describes the 40 Summer Olympic sports. Such dictionaries were inspired by the lexicographic models applied in *FrameNet* (cf. Fontenelle 2003, Ruppenhofer *et al.* 2016) and *Kicktionary* (Schmidt 2007, 2008, 2009), projects which were developed for NLP purposes and also to meet the needs of specialised users – such as language researchers, teachers, and students. The dictionaries by SemanTec, in contrast, were compiled for lay people. Also, in the SemanTec dictionaries novel lexicographic designs were proposed, adapted to the potential users of these tools. As a result, one of the main concerns during the development of the *Field* and *Olympic* dictionaries was the reflection on how the lexicographic model based on frames would serve the target audiences of these tools.

At this point, it is important to note that in a lexicographic project that adopts a frame- based approach, the dictionary functions and the content and structure definitions result not only from the definition of the user's profile, but also from parameters derived from the adopted theoretical model. This approach determines, for example, (i) the users purpose of using the dictionary (encoding and decoding); (ii) the format of the definitions (two-parts definitions: the frame-setting part and the word-specific defining part (Fillmore

[1] Available at: http://dicionariofield.com.br/.

[2] Available at: http://www.dicionarioolimpico.com.br/.

2003: 267)); and (iii) the presentation format (structure) of the dictionary content (relation between a frame and its LUs, relations between frames, and relations between LUs): "a single background frame, entered only once, can serve many word senses, its description could be made accessible from all of the relevant entries" (Fillmore 2003: 263). Alongside these decisions, the definition of the target audience, in turn, complements this process to the extent that it refines the decisions made. When stipulating, for example, a linguists' audience, the choices about function, content and structure take into account the activities in which these professionals engage (reading, writing, listening, speaking, translating); the linguistic skills necessary for the performance of such activities (lexical, morphological, syntactic, semantic, phonological, etymological, metalinguistic knowledge); the degree of familiarity with the medium in which the dictionary will be made available (book, computer, smartphone, internet); and so on.

Thus, it can be said that a dictionary's planning process can be subdivided into three phases: in the first, the characteristics that come from the theoretical model are determined – in this case, the model based on frames; then, the user profile is defined; and, finally, the dictionary's function, content and structure definitions (in that order), which are based on the user profile, are established. Atkins & Rundell (2008), when approaching the pre-lexicographic stage from a more practical perspective, list eight categories according to which the properties of any dictionary should be defined: a dictionary's language(s); a dictionary's coverage; a dictionary's size[3]; a dictionary's medium; a dictionary's organization; the users' language(s); the users' skills; and what users use the dictionary for (cf. Atkins & Rundell 2008: 24-25). The *Field* and *Olympic* dictionaries will now be described using the categories proposed by Atkins & Rundell, but classifying them according to the three phases of a dictionary's planning process. Section 2.1 presents the definitions in terms of the theoretical approach – the dictionary's organization (theoretical perspective); Section 2.2 presents the user-profile definition – the users' language(s) and the users' skills; and Section 2.3 covers the dictionary function, content, and structure definitions – the dictionary's language(s), the dictionary's coverage, the dictionary's medium, the dictionary's organization (practical perspective), and what users use the dictionary for. It should be observed that this section will only present the *Field* and *Olympic* dictionaries; the dictionaries' properties will be problematized in Section 3.1.

2.1 The frame-based lexicographic approach

The theoretical approach adopted to compile a dictionary has implications for decisions regarding the organization of any dictionary. In dictionaries based on the notion of frame, this implies that the description of the lexicon is based on two types of definition, which are complementary and interdependent – the definition of frames and the definition of lexical units; such an organization reflects a conception of meaning based on the continuities between language and experience (i.e., awareness of the physical and social world) (Fillmore 1982, 1985). By adopting this approach for the development of *Field* and *Olympic* dictionaries, the SemanTec group showed interest in representing the meanings of the domains described in a more contextualised and complete way – full knowledge of word meanings (cf. Fillmore (1985), Chishman *et al.* (2014), Chishman *et al.* (2018), Chishman *et al.* (2019)). Thus, from a theoretical perspective of the development of SemanTec's dictionaries, the assumed model implies the adoption of the notions of frame and lexical unit, as evidenced by the excerpts presented in Table 1.

Table 1 - Metalexicographic excerpts related to the organization of the *Field* and *Olympic* dictionaries

Dictionary's organization (theoretical perspective)	
Field	"The lexical units are organised around the notion of semantic frames." (Chishman *et al.* 2014: 26)
Olympic Dictionary	"an electronic lexicographic resource that presents the Olympic sports lexicon, using the notion of frame (or scenario) as an organizing principle." (Chishman *et al.* 2018: 266)

[3] This item does not immediately apply to online dictionaries; pace Gouws & Tarp (2017).

2.2 The target user group definition

The bibliography of lexicography assigns a central and determining role to the pre- lexicographic step of defining a user profile and defends the importance of such an activity even in cases with the potential to serve such a wide audience that reaching even a basic level of detail becomes difficult (Atkins & Rundell 2008; Gouws 2011; Nesi 2013; Lew & de Schryver 2014; Gouws 2018; Tarp & Gouws 2019; Gouws 2020; amongst others). This is because it is from the user-profile definition – and, consequently, from the needs of these users – that many of the lexicographic decisions are made. According to this basic premise, SemanTec's earlier scientific output reveals an intense concern with the user of *Field* and *Olympic* dictionaries, as this output discusses the adaptations and adjustments aimed to serve the target audience better through relevant content and a user-friendly interface. It is worth mentioning, however, that the profile definitions are broad and wide and can be seen as the result of the exercise of placing the intended audience (lay, non-specialised) in an opposite field to that of *FrameNet* and *Kicktionary* users (specialised and NLP). In addition, the themes of these dictionaries (football and Olympic sports) were considered to be of general interest, i.e., of interest to users who do not necessarily have a direct relationship with sports, such as translators, students from many fields, or any person interested in these topics (cf. dos Santos & Chishman 2015; da Silva 2018; Chishman *et al.* 2019). Tables 2 and 3 show excerpts from SemanTec's publications that characterise the user profiles based on the language and skills definitions.

Table 2 - Metalexicographic excerpts related to the users' language(s) of the *Field* and *Olympic* dictionaries

Users' language(s)	
Field	"Portuguese, English, or Spanish first or second language speakers." (dos Santos & Chishman 2015: 449) "It could be accessed by Portuguese, English, or Spanish speakers, either as a first language or additional language." (dos Santos & Chishman 2015: 449)
Olympic Dictionary	Portuguese speakers, either as first language or additional language.

Table 3 - Metalexicographic excerpts related to the user skills of the *Field* and *Olympic* dictionaries

Users' skills	
Field	"A football dictionary aimed at the non-specialised audience." (Chishman *et al.* 2014: 26) "A dictionary for a lay audience." (Chishman *et al.* 2014: 34) "Familiarity with printed and electronic dictionaries, mastery of online tools in different supports (computer, tablet, cell phone)." (dos Santos & Chishman 2015: 449) "The reader is likely to be familiar with other online resources, including other electronic dictionaries, in addition to being sufficiently familiar with the traditional structure of a lexicographic resource." (dos Santos & Chishman 2015: 449)
Olympic Dictionary	"is part of a proposal aimed at a non-specialised and quite heterogeneous audience." (Chishman *et al.* 2018: 273) "user who is not familiar with linguistic theories." (Chishman *et al.* 2018: 272)

With regard to Table 2, it is important to note that the characterization of the users' language(s) in the *Olympic Dictionary* is not described in the texts of the SemanTec group; it was proposed within the scope of the present work based on the characteristics of the language of the *Olympic Dictionary*. From Table 3, it is possible to state that the user profiles are quite broad and wide; both dictionaries specifically (i) refer to lay/non-expert users; while for the *Field* dictionary a user must (ii) be familiar with print and online dictionaries and (iii) have mastered the use of various online tools.

2.3 The dictionary function, content, and structure definitions

Considering that LSP dictionaries are born to meet some demand of a certain public in relation to any description gap of one or more specialised languages, the planning and compilation activities of such dictionaries have to take into account the needs of those who will be the users of these tools (cf. Gouws 2011, 2018). For that matter, user profiles are the starting point for definitions related to (i) the functions of the dictionary – the ways in which the dictionary foresees the linguistic activities with which the target audience engages; (ii) the dictionary's content – the elements that the dictionary offers so that users are able to find answers to their questions; and (iii) the dictionary structure – the path that the dictionary offers so that users can easily and objectively access the answers they are looking for.

At this point, it is worth reminding that the category 'dictionary's organization' was doubled, in order to be able to analyse it from a practical perspective as well. In the case of dictionaries from the SemanTec group, this means analysing the consequences of adopting the notions of frame and lexical unit for structuring the dictionary interface. Moreover, considering that a recurrent theme in SemanTec's work is the adaptation (of content and structure) of the models followed by *FrameNet* and *Kicktionary*, it is relevant to add some extra descriptive information from the dictionaries regarding these adaptations in order to complement the information which was presented based on the categories of Atkins & Rundell (2008). Table 4 presents publication excerpts from SemanTec that characterise the intended uses by the target audiences.

Table 4 - Metalexicographic excerpts related to the use the target user groups make of the *Field* and *Olympic* dictionaries

What users use the dictionary for [Dictionary function]	
Field	"Quick consultation during the 2014 World Cup games – decoding. Use for additional language production – encoding." (dos Santos & Chishman 2015: 449) "We consider that the first type of the dictionary's use would be linked to decoding, when the consultants had to make a quick query to understand a certain term. However, we do not rule out the use of the resource for encoding, in contexts linked, for example, to translation processes." (dos Santos & Chishman 2015: 449)
Olympic Dictionary	Partial encoding and decoding.

In relation to Table 4, it is again worth mentioning that the characterization of the possible uses of the *Olympic Dictionary* by its users is not included in SemanTec publications; therefore, the proposition of the partial encoding and decoding functions is restricted to this work and considers the influence that the methodology adopted for the planning of *Field* had on the planning of the *Olympic Dictionary* and reflects how the dictionary content actually meets these functions. With regard to *Field*, the decoding function seems to be related to quick queries aimed at understanding terms, while encoding is related to translation activities.

Table 5 presents publication excerpts from SemanTec that characterise the languages of the dictionaries.

Table 5 - Metalexicographic excerpts related to the *Field* and *Olympic* dictionaries' languages

Dictionary's language(s) [Dictionary content]	
Field	"Trilingual (Portuguese, English, or Spanish)." (dos Santos & Chishman 2015: 448) Tridirectional "a [...] football dictionary called *Field Dictionary* [...], a trilingual resource (in English, Spanish, and Portuguese)" (Chishman *et al.* 2019: 623)
Olympic Dictionary	Bilingual (Portuguese, English) Unidirectional (Atkins & Rundell 2008) or monodirectional (Welker 2008): the dictionary only presents translation equivalents and examples (both in EN) for the lexical units. "the *Olympic Dictionary* is considered a bilingual resource and unidirectional or [...] monodirectional [...] that is, it is a resource that only allows access to the information that constitutes it in the sense of the source language (Portuguese language) for the target language (English) and not the other way around." (da Silva 2018: 78)

Considering that the definition of the languages in a dictionary is an unfolding of the target user groups' language definition in the dictionary, it is worth highlighting the differences (presented in Table 5) between *Field* and the *Olympic Dictionary* with regard to these properties: although the two dictionaries were designed for very similar purposes, *Field* is a trilingual tridirectional dictionary while the *Olympic Dictionary* is a bilingual unidirectional one.

Table 6 presents the excerpts from SemanTec's publications related to the dictionaries' coverages.

Table 6 - Metalexicographic excerpts related to *Field* and *Olympic* dictionaries' coverages

Dictionary's coverage [Dictionary content]	
Field	"Football specific domain." (dos Santos & Chishman 2015: 448) "a [...] football dictionary called Field Dictionary [...], a trilingual resource (in English, Spanish, and Portuguese)" (Chishman *et al.* 2019: 623)
Olympic Dictionary	"describes the lexicon of [the] 40 Olympic sports." (Chishman *et al.* 2018: 623)

In the year that the *Field* dictionary was launched, Brazil hosted the 2014 FIFA World Cup, and in the year of the launch of the *Olympic Dictionary* Brazil hosted the 2016 Summer Olympics. In this sense, it can be said that the definitions of coverage for both dictionaries were established as an attempt to meet a demand in relation to these events.

Table 7 presents the excerpts from SemanTec's publications related to the dictionaries' mediums.

Table 7 - Metalexicographic excerpts related to the *Field* and *Olympic* dictionaries' mediums

Dictionary's medium [Dictionary structure]	
Field	Digital: online/internet "electronic medium; website with mobile version." (dos Santos & Chishman 2015: 448)
Olympic Dictionary	Digital: online/internet "it is a digital lexicographic product" (da Silva 2018: 52)

With regard to the dictionaries' mediums, the main difference between the *Field* and *Olympic* dictionaries concerns the fact that the *Olympic Dictionary* does not have a smartphone version whose configuration has been designed to adapt the content according to the specifics for devices of this nature.

Table 8 presents the excerpts from SemanTec's publications related to the dictionaries' organizations from a practical perspective.

Table 8 - Metalexicographic excerpts related to the *Field* and *Olympic* dictionaries' organisations (practical perspective)

Dictionary's organization (practical perspective) [Dictionary structure]	
Field	"we propose a macrostructure along the lines of *Kicktionary*. However, we defend the display of two concurrent lists - that of words and that of scenarios" (dos Santos & Chishman 2015: 462)
Olympic Dictionary	"When selecting one of the forms of access, users are directed to one of the three levels of the *Olympic Dictionary*: the modality level, the scenario level, or the word level." (Chishman *et al.* 2019: 626)

Given that this paper started from the premise that in addition to the implications for the organization of the dictionary at a theoretical level, the definitions derived from the frame- based approach must also have consequences for the organization of data in the dictionary application (i.e., website). Thus, in the case of *Field*, for example, the notions of frame and lexical unit guided an organization of the application based on two forms of access (frames and LU lists) and two types of microstructure (the frame and the LU microstructure). In relation to the *Olympic Dictionary*, these characteristics change due to the increase in the number of described sports: to the two forms of access presented by *Field*, *Olympic Dictionary* incorporates a third access format (the sports grid) and, consequently, a third type of microstructure (the sports microstructure).

Table 9 presents the excerpts from SemanTec's publications related to general adaptations applied to *Field* and *Olympic* dictionaries' content and structure.

Table 9 - Metalexicographic excerpts related to the *Field* and *Olympic* dictionaries' adaptations (content and structure)

Adaptations [Dictionary content and structure]	
Field	"The lexical units are organised around the notion of semantic frame, which, in the context of this feature, is called scenario." (Chishman *et al.* 2014: 26) [CONTENT] "Adaptation of the methodological procedures applied in the *FrameNet* platform, since the organization of the information should consist of a friendly interface [STRUCTURE], showing only what is relevant to the reader." (Chishman *et al.* 2014: 34) [CONTENT]
Olympic Dictionary	"(i) adapt information that appears in *FrameNet*, in order to be more easily understood by the layperson, and (ii) suppress information that would not be relevant for this type of user." (Chishman *et al.* 2018: 273) [CONTENT] "the *Olympic Dictionary* maintains the decision taken in the *Field* development process to replace the concepts [frame and lexical unit] with the notions of scenario and word respectively." (Chishman *et al.* 2018: 273) [METALANGUAGE, cf. de Schryver & Joffe (2005)] "Considering that the way in which *FrameNet* displays the relations between frames presupposes a certain familiarity with the theoretical framework of Frame Semantics, the *Olympic Dictionary* presents this information in order to highlight other dimensions of the relationships between scenarios, such as the organization and classification of events, for example." (Chishman *et al.* 2018: 274) [STRUCTURE] "The *Olympic Dictionary* [...] did not intend to provide its target audience with information about verbal valence or syntactic aspects that would not be useful to them." (Chishman *et al.* 2018: 275) [CONTENT]

A last step in describing the *Field* and *Olympic* dictionaries has to do with approaching the adaptations implemented by SemanTec with regard to the development of these tools. Even though SemanTec took inspiration from the *FrameNet* and *Kicktionary* projects for the development of its dictionaries, considering the differences regarding the user-profile characterization, many aspects of the *Field* and *Olympic* dictionaries are described as alternative ways to present information. In addition, the suppression of some information is justified to the extent that they were seen as not being adequate for the target users of the tools.

3 Towards the *Paralympic Dictionary*

Considering the three categories presented in the previous section, the goal in the present section is to problematize the user profiles of the *Field* and *Olympic* dictionaries in order to be able to define the user of the *Paralympic Dictionary*. First the user definitions of the dictionaries already published as well as their implications are analysed (Section 3.1), then the user profile for the *Paralympic Dictionary* is established (Section 3.2).

3.1 What was learned?

With regard to the *Field* and *Olympic* dictionaries, in addition to the definition of the users' languages, the characterization of the user profiles revolves around three features: (i) lay/non- expert users; (ii) familiarity with print and online dictionaries; and (iii) mastering of the use of various online tools. *Field* mentions the three features; the *Olympic* user profile only mentions the first feature.

If, on the one hand, this characterisation can be considered somewhat imprecise (especially as the extent is questionable to which it is able to provide answers about dictionary function, content and structure), on the other hand, it is necessary to keep in mind that the SemanTec projects used not only this broad user-profile definition as a point of departure, but also took into account the lexicographical structures of *FrameNet* and *Kicktionary*. In other words, what this information reveals is that, although the target user characterization has not directly guided the decisions on the dictionaries' function, content and structure, it has always been, to some extent, present (as indicated in the academic writings of the group) in the reflections on the adaptations of elements from *FrameNet* and *Kicktionary*. Therefore, in order to determine if (and if so, which) features of the *Field* and *Olympic* user profiles can or should be mapped onto the *Paralympic* user profile, it is necessary to reflect on two issues:

1) How did the research group integrate the user profile into the analysis of the lexicographical structures of *FrameNet* and *Kicktionary*?

2) Considering this methodology, to what extent was the user profile sufficient to guide the planning of these dictionaries?

Regarding the first question, it can be said that the user profile was integrated in two ways. First, it guided decisions regarding *which* elements (content and structure) should be imported into the group's dictionaries; second, the user profile guided the decisions regarding *how* these elements from the specialised tools should be adapted/reformulated for the interface of the group's dictionaries.

Regarding the second question, the answer is a little more complex. This is because, while the user profile, although broad, relatively efficiently guided decision making on adaptations, it did not serve as a basis for surveying the needs of users (which guide the definition of functions of the dictionary, which in turn guide the definitions of content and structure). Thus, the problem was to think about adaptations without a deeper reflection on functions. By adopting the elements from *FrameNet* and *Kicktionary*, SemanTec got a 'two-for-one deal': the functions of these elements also came with the package, even though they were not compatible with the tools' audiences. Therefore, it would have been very useful to have reflected on the following questions:

1) What functions do these elements serve in *FrameNet*/*Kicktionary*?

2) Which of these functions/elements apply to *Field/Olympic* user needs?

3) What adaptations are needed so that these elements best serve the *Field/Olympic* audience?

4) What other elements can be useful to meet the needs of the *Field/Olympic* audience?

Obviously, when dealing with the need to "(i) adapt information that appears in *FrameNet* […] and (ii) suppress information that would not be relevant for this type of user" (Chishman *et al.* 2018: 273), there is, at the very least, a notion about the activities with which the target audience will *not* engage, revealing that the functions were not completely ignored. However, the use of dictionaries occurs in connection with the types of activities in which users *are* engaged (reception and/or production) (cf. Lew 2012; Nesi 2013). Bearing this in mind and knowing that the user-profile definition usually does not reach a maximum level of specificity – "A user profile seeks to characterise the *typical* user of the dictionary, and the uses to which the dictionary is *likely* to be put" (Atkins & Rundell 2008: 28, emphasis added) – a sufficient profiling for these tools should be able to list the most common needs and uses of prototypical users. There is an attempt in this direction in the description of *Field*'s function and user skills: "when the consultants had to make a quick query to understand a certain term" (dos Santos & Chishman 2015: 449) and "Familiarity with printed and electronic dictionaries, mastery of online tools in different supports (computer, tablet, cell phone)" (dos Santos & Chishman 2015: 449). However, these are questionable points since (i) the dictionaries do not have word definitions and, therefore, the user cannot quickly resolve doubts about terms; (ii) the dictionaries do not follow a traditional orientation, so being familiar with the structure of traditional dictionaries does not necessarily improve the experience; and, finally, (iii) it is not clear to what extent the structures of the dictionaries are based on the structures of other online tools (such as social networks, for example) to the point where it is possible to say that the experience that users bring from other sites can contribute to a better experience in using SemanTec's dictionaries.

3.2 What is next?

Regarding the *Paralympic Dictionary* user-profile definition, the discussion resonates in ways to encourage a more detailed user profiling but also to reinforce the relevance of considering such a profile when establishing the functions, content and structure of a dictionary. Following the same approach employed for the analysis of the *Field* and *Olympic* user profiles, in this section the categories proposed by Atkins & Rundell (2008: 24-25) are again the guiding principle.

Table 10 - Paralympic user-profile definition

Users' language(s)	➢ Portuguese and English native speakers, either as first language or as additional language.

Target users	➢ Linguists and other language professionals;
	➢ Mass media professionals;
	➢ Literate adults;
	➢ School students;
	➢ Middle childhood;
	➢ Language learners;
	➢ Athletes with disabilities and other sports' professionals;
	➢ People with disabilities (such as vision impairment, deafness, dyslexia, etc.);
	➢ People familiar with a variety of online resources (smartphone, tablet, computer).

Once one has outlined the user profile, the needs of each of these groups may be listed and, based on these, the functions, content and structure of the dictionary may be defined. Regarding the needs of language learners, for example, it is possible to list needs related to reading and writing, but also listening and speaking. Such demands guide the definition of functions aimed at serving this specific group (encoding and decoding), the proposition of content elements (definition of words, definition of frames, examples in the target language, translation equivalents, etc.) and, finally, the definition of the dictionary structure (dictionary portal structure (cf. Gouws 2018); with three levels of information: superframe, frame, and word), etc. From the people with disabilities' perspective, definitions of structure also influence decisions with deeper implications that promote inclusion through digital accessibility (cf. Chishman *et al.* 2021). Finally, it is worth mentioning that the survey of the needs of the different user groups who are part of this profile will demand a more detailed investigation of the activities with which these people are involved and in what sense the dictionary can contribute to the performance of such activities.

4 Conclusion

The development of the *Paralympic Dictionary* is an extension to the dictionaries already published by the SemanTec research group. In addition to new issues (such as discussions involving inclusion, both from the perspective of access to the dictionary and the sports' presentation and representativeness), the work with Paralympic sports has motivated reflections aimed at ensuring a better experience of using dictionaries by means of the improvement of theoretical and methodological aspects that characterised the previous projects. Among the theoretical-practical issues, the definition of the user profile was emphasised, given its centrality in the processes of dictionaries' planning and production. In this sense, it was considered relevant to reassess the user-profile definitions of the *Field* and *Olympic* dictionaries in order to determine the starting point for a reflection of this nature in the *Paralympic* context, listing the positive and negative aspects to be taken into account and pointing out perspectives for the next steps.

The analysis revealed that the concepts adopted by the SemanTec team members in relation to the target audience of the dictionaries already compiled resulted in tools that were relatively distant/disconnected from the specific situations in which they are (or could be) used. This is because the elements that constitute the dictionaries do not necessarily represent ways to meet the specific needs of users, but adaptations based on a conception of abstract users with abstract needs ('massification' – cf. Tarp & Gouws (2019)). This finding does not suggest that the tools are not used by real users with real needs, but from the point of view of the properties that characterise the dictionaries, it is difficult to specify who these users actually are. Therefore, this discussion is extremely relevant for the definition of the *Paralympic* user to avoid these generalizations and to meet the needs of its target audience.

5 References

Atkins, B.T. Sue & Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York, NY: Oxford University Press.

Chishman, Rove. 2014. *Field - Dicionário de Expressões do Futebol. 2.ed. [Field – Football Expressions Dictionary. 2nd ed.]*. São Leopoldo: Unisinos.

Chishman, Rove. 2016. *Dicionário Olímpico [Olympic Dictionary]*. São Leopoldo: Unisinos. Chishman, Rove, Aline Nardes dos Santos, Diego Spader de Souza & João Gabriel Padilha. 2014. Field – Dicionário de Expressões do Futebol: um recurso lexicográfico baseado no aporte teórico-metodológico da semântica de frames e da linguística de corpus [Field – Football Expressions Dictionary: A lexicographic resource based on the theoretical- methodological approach of Frame Semantics and Corpus Linguistics]. *Signo* 39(67): 25– 35.

Chishman, Rove, Larissa Moreira Brangel, Diego Spader de Souza, Aline Nardes dos Santos, Bruna da Silva & Sandra de Oliveira. 2018. *Dicionário Olímpico*: a semântica de frames encontra a lexicografia eletrônica [*Olympic Dictionary*: Frame Semantics meets electronic lexicography]. In: Finatto, Maria José Bocorny, Rozane Rodrigues Rebechi, Simone Sarmento & Ana Eliza Pereira Bocorny (eds). *Linguística de corpus: perspectivas*: 265–98. Porto Alegre: Instituto de Letras - UFRGS.

Chishman, Rove, Aline Nardes dos Santos, Bruna da Silva & Larissa Brangel. 2019. Challenges and difficulties in the development of *Dicionário Olímpico* (2016). In: Kosem, Iztok, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek & Carole Tiberius (eds). *Electronic Lexicography in the 21st Century (eLex 2019): Smart lexicography. Conference proceedings. Sintra, Portugal, 1-3 October 2019*: 622–41. Brno: Lexical Computing.

Chishman, Rove, Gilles-Maurice de Schryver, Bruna da Silva, Aline Nardes dos Santos, Ana Luiza Treichel Vianna, Sandra de Oliveira & Mikaela Luzia Martins. 2021. Building a Paralympic, frame-based dictionary – Towards an inclusive design for *Dicionário Paraolímpico* (Unisinos/ Brazil). In: Gavriilidou, Zoe, Maria Mitsiaki & Asimakis Fliatouras (eds). *Proceedings of the XIXth EURALEX Congress: Lexicography for Inclusion, Vol. II*: 8 pages. Alexandroupolis: Democritus University of Thrace.

da Silva, Bruna. 2018. *Lexicografia eletrônica e semântica de frames: o potencial da noção de frame para o desenvolvimento de dicionários digitais online [Electronic lexicography and Frame Semantics: The potential of the notion of frame for the development of online digital dictionaries]*. MA dissertation. São Leopoldo: Universidade do Vale do Rio dos Sinos.

de Schryver, Gilles-Maurice & David Joffe. 2005. Dynamic metalanguage customisation with the dictionary application TshwaneLex. In: Kiefer, Ferenc, Gábor Kiss & Júlia Pajzs (eds). *Papers in Computational Lexicography, COMPLEX 2005*: 190–99. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

dos Santos, Aline Nardes & Rove Chishman. 2015. O papel da semântica de frames na construção de um recurso dicionarístico: a organização lexicográfica do Field – Dicionário de Expressões do Futebol [The role of Frame Semantics in the construction of a dictionary: The lexicographical organization of Field – Football Expressions Dictionary]. *Revista da ABRALIN* 14(3): 433–68.

Fillmore, Charles J. 1977. Scenes-and-frames semantics. In: Zampolli, Alan (ed.). *Linguistic Structures Processing* (Fundamental Studies in Computer Science 59): 55–88. Dordrecht: North Holland Publishing.

Fillmore, Charles J. 1982. Frame Semantics. In: The Linguistic Society of Korea (ed.).*Linguistics in the Morning Calm*: 111–37. Seoul: Hanshin Publishing Co.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2): 222–54.

Fillmore, Charles J. 2003. Double-decker definitions: The role of frames in meaning explanations. *Sign Language Studies* 3(3 - Special Issue on 'Dictionaries and Lexicography, Part I: General Issues in Lexicography'): 263–95.

Fontenelle, Thierry (ed.). 2003. *Special Issue on FrameNet and Frame Semantics* (International Journal of Lexicography 16.3). Oxford: Oxford Journals.

Gouws, Rufus H. 2011. Learning, unlearning and innovation in the planning of electronic dictionaries. In: Fuertes-Olivera, Pedro A. & Henning Bergenholtz (eds). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 17–29. Londen: Continuum.

Gouws, Rufus H. 2018. Internet lexicography in the 21st century. In: Engelberg, Stefan, Heidrun Kämper & Petra Storjohann (eds). *Wortschatz: Theorie, Empirie, Dokumentation* (Germanistische Sprachwissenschaft um 2020 2): 215–36. Berlin: De Gruyter.

Gouws, Rufus H. 2020. Special field and subject field lexicography contributing to lexicography. *Lexikos* 30: 143–70.

Gouws, Rufus H. & Sven Tarp. 2017. Information overload and data overload in lexicography. *International Journal of Lexicography* 30(4): 389–415.

Lew, Robert. 2012. How can we make electronic dictionaries more effective? In: Granger, Sylviane & Magali Paquot (eds). *Electronic Lexicography*: 343–61. Oxford: Oxford University Press.

Lew, Robert & Gilles-Maurice de Schryver. 2014. Dictionary users in the digital revolution. *International Journal of Lexicography* 27(4): 341–59.

Nesi, Hilary. 2013. Researching users and uses of dictionaries. In: Jackson, Howard (ed.). *The Bloomsbury Companion to Lexicography*: 62–74. London: Bloomsbury.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, Collin F. Baker & Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Schmidt, Thomas. 2007. The Kicktionary: A multilingual resource of the language of football. In: Rehm, Georg, Andreas Witt & Lothar Lemnitzer (eds). *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference*: 189–96. Tübingen: Gunter Narr.

Schmidt, Thomas. 2008. The Kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary. In: Lavric, Eva, Gerhard Pisek, Andrew Skinner & Wolfgang Stadler (eds). *The Linguistics of Football* (Language in Performance 38): 11–21. Tübingen: Gunter Narr.

Schmidt, Thomas. 2009. The Kicktionary – A multilingual lexical resource of football language. In: Boas, Hans C. (ed.). *Multilingual FrameNets in Computational Lexicography*: 101–34. Berlin: De Gruyter Mouton.

Tarp, Sven & Rufus H. Gouws. 2019. Lexicographical contextualization and personalization: A new perspective. *Lexikos* 29: 250–68.

Welker, Herbert A. 2008. Lexicografia pedagógica: Definições, hist**ó**ria, peculiaridades. In: Xatara, Claudia, Cleci Bevilacqua & Philippe Humblé (eds). *Lexicografia Pedagógica: Pesquisas e Perspectivas*: 9–45. Santa Catarina: NUT – Núcleo de Tradução, Ed. Universidade Federal de Santa Catarina.

PROVIDING ETYMOLOGICAL INFORMATION FOR SINITIC LOANWORDS IN THE KBBI INDONESIAN DICTIONARY

David Moeljadi
Kanda University of International Studies, Japan
davidmoeljadi@gmail.com

Abstract

This paper documents the process of adding the etymological information of loanwords from Sinitic languages in Indonesian language into the KBBI Indonesian dictionary fifth edition, the most comprehensive and authoritative Indonesian monolingual dictionary, published by The Language Development and Cultivation Agency, under the Ministry of Education and Culture. It is a part of the etymology project which involves experts from universities in Indonesia (Moeljadi et al. 2019). Data of Sinitic loanwords from various sources such as Schlegel (1891), Hamilton (1924), Png (1967), Leo (1976), Kong (1994), and Jones (2009) were compiled. Data selection is based on the dictionary headwords, thus words which are listed in the KBBI dictionary were chosen and further analyzed. Finally, a database of Sinitic loanwords for the KBBI dictionary was built.

Historically, there are four major periods associated with external cultural and linguistic influence in the Indonesian archipelago: (1) Indian, (2) Chinese, (3) Islamic, and (4) European (Blust 2009). As of February 2021, the KBBI dictionary has etymological information of loanwords from Semitic languages (especially Arabic) and Indic languages (especially Sanskrit). Since languages in southern part of China were the early donor languages, it is worth adding the etymological information of loanwords from those languages into the KBBI dictionary.

The earliest instance of a Sinitic loanword is *tahu* ‘bean curd’ which is attested in an Old Javanese inscription from the tenth century (Jones 2009). Some tools such as *gunting* ‘scissors’ which is also found in Old Javanese texts might be borrowed from a southern Chinese language (Blust 2009). In the early Ming period after 1368, various Sinitic loanwords were borrowed through trade such as *opau* ‘money belt, small wallet’ and *honcoe* ‘smoking pipe’ (Blust 2009).

I found that there are more than 350 Sinitic loanwords in the KBBI dictionary. Regarding semantic domains, many of them are related to food, tradition and customs, and commerce, the rests are related to tools, clothes, kinship terms, martial arts, opium, prostitution, medicine, etc. Regarding donor languages, most of them are from Hokkien, others are from Cantonese, Hakka, and Mandarin.

Keywords: Sinitic loanwords, KBBI dictionary, lexical borrowing

1 Introduction

Kamus Besar Bahasa Indonesia (KBBI) is the official dictionary of the Indonesian language, published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency) or Badan Bahasa, under the Ministry of Education and Culture, Republic of Indonesia. Up until present, KBBI is the most comprehensive and the most authoritative reference for the Indonesian language. Etymological information was added in October 2019 for Semitic (especially Arabic) loanwords and in October 2020 for Indic (especially Sanskrit) loanwords. This paper discusses the inclusion of etymological information from Sinitic languages into the KBBI database which is planned in October 2021. It is a part of the KBBI etymology project (Moeljadi et al. 2019).

1.1 Scope of research

The term "Sinitic loanwords" in this paper refers to those loanwords from various languages in China, especially Hokkien, Hakka, Cantonese, and Mandarin, which are completely borrowed and thus listed as words in Kamus Besar Bahasa Indonesia (KBBI). Sinitic loanwords appear to be more widely used in the late colonial period than in more recent times, both by Chinese and non-Chinese speakers (Hoogervorst 2017). For example, the personal pronouns *gua* (我) 'I' and *lu* (汝) 'you' have been taken over by non-Chinese speakers of Betawi Malay and several other varieties. These pronouns are still used nowadays. The use of pronouns *bwansing* (晚生) 'I' and *owe* (喂) 'I' is restricted to the ethnic Chinese (Nio 1955: 43-44). To the best of my knowledge, these pronouns can be considered as archaic.

It is important to make a distinction between Sinitic lexical influence that has entered the mainstream Malay/Indonesian language and loanwords only understood by ethnic Chinese (Leo 1975). Thus, there are two types of Sinitic loanwords. The first one is those which are used by both ethnic Chinese and non-Chinese speakers, standardized, and thus listed in KBBI.[1] The second one is those which are used only by ethnic Chinese, not standardized (there are variations in spellings) and thus not listed in KBBI. Regarding the first type, we can divide into two groups: the first one is those which are used until present-time or those having attested status as words in the present language, whether people are aware that they are borrowed or not, for example *mi* 'noodles' and *tahu* 'tofu, bean curd'. The second one is archaic words, such as *kimantu* 'derogatory term for a Chinese newcomer'. Such words are labelled "ark" (a short form of *arkais* 'archaic') in KBBI. Similarly, regarding the second type, we can divide into two groups, i.e. those which are used until present-time (widely or narrowly in some communities) such as *Cungkuo* 'China' and those which were used in the past time or archaic, such as *owe* 'I'. The present paper only deals with the first type, i.e. those which are listed in KBBI. See Figure 1 for the types of Sinitic loanwords.

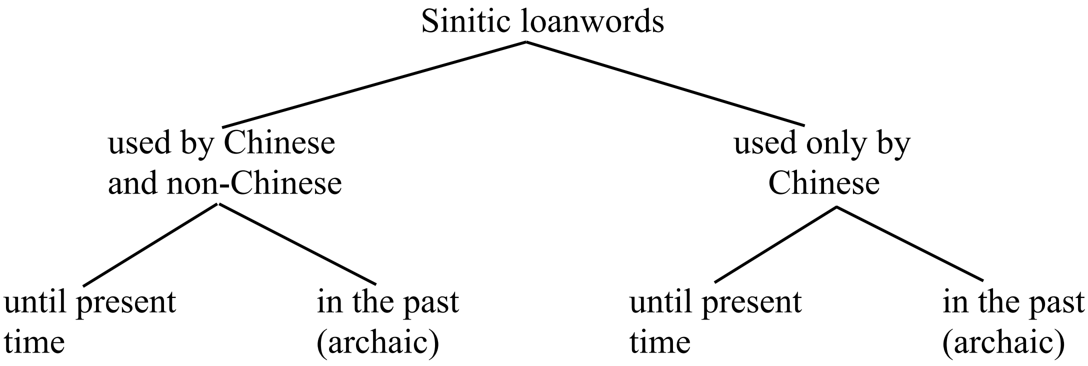


Figure 1. Types of Sinitic loanwords in Indonesian

In addition to Sinitic loanwords, there are Sinitic loan translations, as well as hybrid forms and ad hoc creations. Constructions such as *nasi pagi* 'breakfast' or *makan pagi* 'breakfast', *moeloet pintoe* 'doorway' and *keloear pintoe* 'to go out' are literal translations of Chinese 早飯, 早膳, 門口, dan 出門 respectively (Salmon 1974). The present paper does not deal with these loan translations although some of them are listed in KBBI.

1.2 Historical background and Sinitic influence in Malay lexical sources

Contact with speakers from China had happened from the seventh to the tenth century A.D., when Chinese merchants traded to Riau Islands, West Kalimantan, and East Kalimantan, even until North Maluku, long

[1] KBBI has language labels. For Chinese or Sinitic languages, the language label is "Cn" (a short form of *Cina* 'Chinese'). Lexical entries which are used particularly by ethnic Chinese are given the label "Cn".

time before the arrival of Portuguese and Dutch people. When Sriwijaya kingdom appeared and became strong, China also opened a diplomatic relation with Sriwijaya to secure its trade and shipping business. In the year 922, Chinese travelers visited Kahuripan kingdom in East Java. Since the 11th century, hundreds of thousands of Chinese migrants left their ancestral land and settled in many parts of the Archipelago.[2] During the Dutch colonial period, more Chinese migrants who were contracted by the Dutch came to the archipelago. The Chinese population increased. In 20th century during the revolutionary movement in China, more and more Chinese people came to Indonesia.

The influence of Sinitic languages in Malay can be seen from the lexical sources. Before paper dictionaries, there are word-lists or lists of words in a foreign language with the equivalent meaning in Malay. The earliest extant word-list of Malay is a Chinese-Malay vocabulary, dated to the 15th century, containing 482 entries which is written wholly in Chinese characters and employed both to give the Chinese word and for a transcription of the sound of the Malay word (Edwards and Blagden 1931). It already contains a number of Sinitic loanwords. The next word-list is the one compiled by Antonio Pigafetta, an Italian, from materials collected in about 1521, probably from the eastern islands of Indonesia. This contains some 426 items, the Italian word being given first, followed by the Malay equivalent (Marsden 1984). It has at least one Sinitic loanword. From colonial times, all documented varieties of Malay seem to have undergone some degree of Chinese influence. During the 19th century and the beginning of 20th century of colonial period, some word-lists related to Sinitic loanwords were published in the Archipelago, such as Schlegel (1891) and Hamilton (1924).

In the 20th century, there are a number of publications related to Sinitic loanwords which are mentioned in the following chapter.

2 Method

Data of Sinitic loanwords in Malay/Indonesian from various sources were gathered and compiled from 2019 to 2020. Table 1 summarizes the data sources.

Table 1. Data sources

No.	Source	Number of Sinitic loanwords	Comments
1	Blust and Trussel (2010)	8	The Austronesian Comparative Dictionary (web edition) with information on loanwords[3]
2	Chow (2010)	515	M.A. thesis on Sinitic loanwords in Standard Malay
3	Hamilton (1924)	189	Sinitic loanwords in Malay Peninsula
4	Jones (2009)	1,469	Sinitic loanwords in Indonesian and Standard Malay
5	Kong (1994)	1,046	Sinitic loanwords in Indonesian and Standard Malay
6	Leo (1976)	288	Sinitic loanwords spoken by the inhabitants of Jakarta
7	Png (1967)	416	Sinitic loanwords in Malay
8	Schlegel (1891)	92	Sinitic loanwords in Malay
9	Sutami (2016)	407	Sinitic loanwords in Indonesian

In addition, eleven words which are not mentioned in any of those sources but are listed in KBBI were

[2] The Archipelago or *Nusantara* refers to Malay-related cultural and linguistic lands, such as the present Indonesia, Singapore, Brunei, and Malaysia. Standard Malay spoken in Malaysia and Indonesian spoken in Indonesia are two standardized varieties of the Malay language. There are other Malay varieties such as Singapore Malay and Brunei Malay.

[3] https://www.trussel2.com/acd/acd-lo_a.htm

manually added, i.e. *butongpai* ‘a kind of martial arts’, *micin* ‘MSG, vetsin’, *laucu* ‘Laozi’, *tokwi* ‘tablecloth’, *shou sui* ‘a tradition on the night of Chinese New Year’, *syantung* ‘finely woven cloth from Shandong’, *takoah* ‘bean curd skin’, *hoisem* ‘sea cucumber’, *saucu* ‘grilled pork’, *ako* ‘elder brother’, and *dizi* ‘Chinese flute’.

Loanwords from each source, together with its details and explanations, were gathered and summarized in a table which contains the following information: part-of-speech, ID, Indonesian word in KBBI, Indonesian word in the source, Indonesian word meaning, source language, original word, Chinese character, original meaning, semantic domain, and earliest in corpus. The Indonesian words and parts-of-speech are taken from KBBI. The Indonesian/Malay words, word meanings, source languages, original words, Chinese characters, and original meanings are taken from the source directly. Thus, the Indonesian/Malay words in the sources and Indonesian words in KBBI sometimes differ in orthography. The ones in KBBI are the standard ones. The data sources also differ in the information on donor languages; some sources have “Chinese”, while some have “Hokkien” or “Hokkian”. The semantic domains were decided by myself. The information on “earliest appearance in corpus” was based on the Malay Concordance Project (MCP)[4] and the Austronesian Comparative Dictionary (ACD). Table 2 contains three example items from the Sinitic loanwords data I have compiled.

Table 2. Some examples of Sinitic loanwords data

ID	SCHLEGEL_83	KONG_364	LEO_2_12
Indonesian word in KBBI	*cuki*	*honcoe*	*ceki*
part-of-speech	noun	noun	noun
Indonesian/Malay word	*tjuki*	*huncue*	*ceki*
Indonesian/Malay word meaning	a kind of draughts played with white and black beans	*paip penghisap tembakau* ‘pipe for smoking tobacco’	a Chinese card game the same as *capjiki*
Source language	Chinese	Fujian/Hokkian	Hokkien
Original word	*tioh kî*	*hun ts'ui*	*tsit-kī*
Chinese character	着棋	熏喙	一枝
Original meaning	to play at draughts or chess	-	*tsit*, one; *kī*, piece; literally, one piece
Semantic domain	gambling	tobacco	gambling
Earliest in corpus	Hikayat Bayan Budiman, 1371 (MCP)	probably during the Ming dynasty (ACD)	Warkah Buton, 1790s? (MCP)

During this compilation process, I found challenges regarding data sources, as follows.

1. Different sources have different information on donor or source languages, as noted in Table 2. Most of the sources only mention Hokkien or Hokkian or Fujian as the donor language while Jones (2009) writes the dialect names of Hokkien, e.g. Amoy (Xiamen) dialect, Changchiu (Zhangzhou) dialect, Tsoanchiu (Quanzhou) dialect etc.

2. Different sources have different romanization systems.

3. Different sources have different Chinese characters, for example the Chinese characters for *cincau* ‘grassjelly’ are 清草 in Jones (2009) and Sutami (2016), 青草 in Kong (1994) and Leo (1976), and 蘔草 in Schlegel (1891). There is no standardization in writing Chinese characters for Hokkien because Hokkien is mainly a spoken language.

4. Authors have different opinions on which words can be considered as Sinitic loanwords, i.e. loanwords which are used in the present modern Indonesian or those which are used in pre-independence or during late-colonialism (Sino-Malay literature).

[4] http://mcp.anu.edu.au/Q/mcp.html

5. There is lack of documentation regarding when a word is recorded for the first time and thus we do not know the first appearance of most of the words in documents and published materials as well as their meaning development.

After the data compilation, I did a manual entry selection process i.e. I chose one entry from the candidate entries based on the amount of information. I chose the most informative and specific one, as illustrated in Table 3. If there is only one entry from one source, I examined and decided if it is a suitable entry and can be selected or it is a doubtful entry and marked it as "doubtful".

Table 3. Entry selection

ID	JONES_544	LEO_1_27	SUTAMI_187
Indonesian word in KBBI	*jok*	*jok*	*jok*
part-of-speech	noun	noun	noun
Indonesian/Malay word	*jok*	*jok*	*jok*
Indonesian/Malay word meaning	seat of car or pedicab*; cf *loanjok*	mattress, seat of car or pedicab	*alas tempat duduk* 'seat cover'
Source language	Chiangchiu (Zhangzhou)	Hokkien	Chinese
Original word	*jio̍k*	*dziók*	*rù*
Chinese character	褥	褥	褥
Original meaning	-	*dziók*, mattress	-
Semantic domain	vehicle	vehicle	vehicle
Notes	selected	-	-

Afterwards, I summarized all the chosen entries in a table for KBBI database. There is a column for data sources. All sources which support the chosen entries should be mentioned in this column. I follow Carstairs Douglas' romanization system for Hokkien (Douglas 1899). The format of the table is shown in Table 4.

Table 4. The chosen entry for *jok* 'seat of car or pedicab' in KBBI database format

node_ID	**entry_ID**	**entry**	**language**	**orig_word**	**translit**	**orig_meaning**	**source_ID**
10199	34790	*jok*	Hokkien Zhangzhou dialect	褥	*jio̍k*	*matras, kasur* 'mattress'	JONES, LEO, SUTAMI

3 Result

After the selection process, there are 357 Sinitic loanwords and 40 doubtful entries. The 357 loanwords belong to six parts-of-speech: 307 nouns, 22 adjectives, 12 numerals, 10 verbs, 4 pronouns, and 2 particles (interjection and expression). It is interesting to know that there are more numerals borrowed than verbs. The borrowed numerals are, for example, *ceban* 'ten thousand', *ceceng* 'one thousand', *cepek* 'one hundred', *goban* 'fifty thousand', *goceng* 'five thousand', and *gopek* 'five hundred' which I think correspond to the Indonesian currency denominations and they are usually used in business or commerce.

Regarding the archaic words, 18 out of 357 Sinitic loanwords (5%) are labelled as archaic in KBBI. Regarding the words mainly used by ethnic Chinese, 88 of them (24.6%) have Chinese as the language label, i.e. people still recognize them as Sinitic loanwords. It means that more than half or most of the Sinitic loanwords are used by both ethnic Chinese and non-Chinese. 3.4% or 12 of them have Jakarta Malay as the language label, i.e. the Sinitic words were borrowed via Jakarta Malay speakers in Jakarta.

In addition, there are two words labelled “Mal” (Standard Malay), one word labelled “Jw” (Javanese), one word labelled “Sd” (Sundanese), and one word labelled “Jb” (Jambi Malay). Through these language labels, we know that some Sinitic words were borrowed via Malay, Javanese, Sundanese, and Jambi Malay.

I classified the Sinitic loanwords into 37 semantic categories as follows: 85 of them (23.8%) are related to food, 48 of them (13.4%) tradition and customs, and 36 of them (10.1%) commerce, the rests are related to clothes (5.9%), kinship terms (3.9%), gambling (3.9%), medicine (3.9%), house (3.9%), tools (3.6%), martial arts (3.6%), seafaring (2.5%), body parts (2.5%), prostitution (2.2%), opium (2%), place (2%), vehicle (1.4%), government (1.1%), tobacco (1.1%), plants (0.8%), animals (0.8%), and others (6.2%). Regarding donor languages, most of them are from Hokkien (89.4%), others are from Cantonese (5.3%), Mandarin (3.1%), and Hakka (1.7%).

4 Analysis and Discussion

This paper has reached its goal in terms of adding etymological information for Sinitic loanwords into the KBBI. However, in order to make a proper, comprehensive etymological information for Sinitic loanwords, etymological information for words particularly used by ethnic Chinese and Sino-Malay words should be added and extensive etymological research based on corpora should be conducted. In order to get more reliable data with years of occurences, digital corpora of Malay/Indonesian should be made. The following data sources can be employed for that purpose.

1. Malay Concordance Project (MCP). I employed this source to check the occurences of Sinitic loanwords in Classical Malay texts between 1302 and 1953. This source contains 165 texts and 5.8 million words, including 140,000 verses. The corpora have been digitized so we can do various kinds of searches.
2. Monash University’s Sin Po newspaper collection.[5] This collection contains newspaper articles in Sino-Malay language between April 1923 and December 1941. It only has PDF and microfilms.
3. University of Washington collection which contains two kinds of Sino-Malay literature, written between 1886 and 2000, owned by two temple libraries in Java.[6] The first one is 5,000 scanned pages from a collection of religious books and magazines in Chinese and Malay languages. The second one is 12,500 pages of popular Sino-Malay novels and magazines.
4. National Library of Indonesia collection of Sino-Malay literature between 1804 and 1950.[7] It contains 14,162 pages from 122 titles of novels, magazines, story books, poems, and picture books. All of them are in PDF.
5. Leiden University Library collection of Sino-Malay literature between 1870s and 1950s.[8] It consists of two newspapers (Sin Po and Hoakiao) and circa 1,400 books (dime novels, educational works, translations, works on religion, poems, and literary works). However, it is not open to people outside Leiden University.

Sources 2 to 5 have the data saved in PDF and/or microfilms. The data should be digitized first before being able to be employed for various kinds of searches. In the end, we should answer the following etymological questions which we cannot answer all of them only by data gathering, compilation, and selection described in this paper:

1. What are the source or donor languages?
2. When were the words borrowed or when did they appear for the first time in which text and what

[5] https://repository.monash.edu/collections/show/117

[6] https://digital.lib.washington.edu/researchworks/handle/1773/21474

[7] http://e-resources.perpusnas.go.id/library.php?id=00031

[8] https://digitalcollections.universiteitleiden.nl/sinomalaytexts

did it mean?

3. Were they directly or indirectly borrowed? What were the intermediate stages of the borrowing process?
4. How did they reflect the social history at that time?
5. Were they borrowed together with many other (semantically related) words?
6. Did they replace the (previous) Malay word?
7. Did they undergo meaning shifts and sound changes?

In addition, Old Javanese and Sinitic sources should be referred to if possible. Regarding Sinitic sources, some of the challenges are to make corpora of Hokkien/Min Nan, Cantonese, and Hakka and to conduct research on etymological information of Sinitic words.

Acknowledgments

I thank Anna Šlajová who helped me input most of the data into an Excel spreadsheet. I gratefully acknowledge the support of the European Regional Development Fund Project "Sinophone Borderlands – Interaction at the Edges" CZ.02.1.01/0.0/0.0/16_019/0000791.

5 References

Blust, Robert A. (2009). *The Austronesian languages*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Blust, Robert and Stephen Trussel. (2010). Austronesian Comparative Dictionary, web edition (www.trussel2.com/ACD). Revision 21/06/2020. Accessed on 22/05/2021.

Chow, Chai Khim. (2010). *The Study of Loanwords between Chinese Language and Malay Language in Malaysia*. M.A. thesis. National University of Singapore and Peking University.

Douglas, Carstairs. (1899). *Chinese-English Dictionary of the Vernacular or Spoken Language of Amoy, with the Principal Variations of the Chang-chew and Chin-chew Dialects*. London: Publishing Office of the Presbyterian Church of England.

Edwards, E. D. and Blagden, C. O. (1931). A Chinese vocabulary of Malacca Malay words and phrases collected between AD 1403 and 1511 (?). *Bulletin of the School of Oriental Studies*, University of London, 6(3), 715-749.

Hamilton, A. W. (1924). Chinese loan-words in Malay. *Journal of the Malayan Branch of the Royal Asiatic Society*, 2(1 (90), 48-56.

Hoogervorst, T. G. (2017). What kind of language was 'Chinese Malay'in late colonial Java?. *Indonesia and the Malay World*, 45(133), 294-314.

Jones, Russell. (2009). *Chinese Loan-words in Malay and Indonesian: A Background Study*. Kuala Lumpur: University of Malaya.

Kong Yuanzhi. (1994). Kata pinjaman Bahasa Cina dalam Bahasa Melayu. In *Jurnal Dewan Bahasa*, Aug., pp. 676-702; Sept., pp. 772-795.

Leo, Philip. (1975). *Chinese Loanwords Spoken by the Inhabitants of the City of Jakarta*. Jakarta: Lembaga Research Kebudayaan Nasional L.I.P.I.

Marsden, William. (1984). *A dictionary and grammar of the Malay language*, [Facsimile reprint of the 1812 ed.], Singapore: Oxford University Press, 2 vols.

Moeljadi, David, Ian Kamajaya, and Azhari Dasman Darnis. (2019). Considerations for Providing Etymological Information in the KBBI Indonesian Dictionary. In Mehmet Gürlek, Ahmet Naim Çiçekler, and Yasin Taşdemir (Eds.), *Proceedings of the 13th International Conference of the Asian Association for Lexicography*, Istanbul University, pp. 161–178. Istanbul: Asos Publisher.

Nio, Joe Lan. (1955). Kata-kata 'goea' dan 'loe' di Indonesia [The words 'goea' and 'loe' in Indonesia]. *Bahasa dan Budaja* 3 (3):41–44.

Png, Poh-Seng. (1967). A preliminary survey of Chinese Loan-words in the Malay Language. The Journal of Southeast Asian History. Singapore: The Island Society.

Salmon, Claudine. (1974). Les traductions de romans chinois en malais (1880–1930) [Translations of Chinese novels into Malay]. In P-B. Lafont and D. Lombard (eds), *Littératures contemporaines de l'Asie du sud-est* [Contemporary literatures of Southeast Asia]. Paris: l'Asiathèque, pp. 183–201.

Schlegel, G. (1891). Chinese Loan-words in the Malay Language. In *T'oung Pao*, I, pp. 391-405.

Sutami, Hermina. (2016). Menelusuri Penggunaan Sumbangan Kosakata Bahasa Cina dalam Bahasa Indonesia. Lampiran tulisan untuk memperingati hari jadi ke-85 Prof. Dr. Muhajir, Guru Besar

DEFINITION MODEL FOR PLANT AND ANIMAL NAME LEMMAS IN KBBI V: A USER STUDY

Dewi Khairiah
National Agency for Language Development and Cultivation, Indonesia
dewi.khairiah@kemdikbud.go.id

Abstract

MacMillan (1949) says that dictionary should be evaluated regarding to the quantity of the information, the quality of the information, and the effectiveness of the information presented in the dictionary, while the quality of the defnition can be examined from its accuracy, completeness, clearness, simplicity, and modernity. Further, the evaluation of dictionary is part of dictionary criticism that is aimed to give inputs for the improvement and development of dictionary (Gouws in Bielinska, 2017). In this user study, 120 defnitions for 30 lemmas of plant and animal names are taken from four sources: Kamus Besar Bahasa Indonesia, Fifth Edition (KBBI V); Kamus Umum Bahasa Indonesia (KUBI); terms dictionaries; and defnition model proposed by researcher. This model is developed from the model entry of Atkinson and Rundell (2008), contains the categorized necessary information in sequence order: taxonomy/scientifc name-characteristics-habitat-additional information. Then, the respondents are asked to rank the information categories they consider as relevant and signifcant as the defniens in defnition. At last, the respondents are asked about their general opinion on the defnitions in KBBI V.

The questionnaire shows three results. First, the respondents consider the information category of taxonomy/scientifc name as the most relevant and signifcant information (100%), followed by the information category of characteristics (97%). Meanwhile, they consider the information category of habitat is as relevant and signifcant as the category of additional information (83%). Three highest preferred formation patterns of defniens are 1) taxonomy/scientifc name-characteristics-habitat-additional information, 2) taxonomy/scientifc name-characteristics-additional information-habitat, and 3) taxonomy/scientifc name-additional information-characteristics-habitat. Second, 39% respondents prefer the proposed defnition model to KBBI's (29%). Third, the evaluation results that KBBI V meet three criteria, they are related to the quantity of the information (92%), the quality of the information (93%), and how the information presented in dictionary (92%). Nevertheless, the issue of circular defnitions in KBBI V still concerns the respondents.

Keywords dictionary criticism, user study, *definiens*, definition

1 Introduction

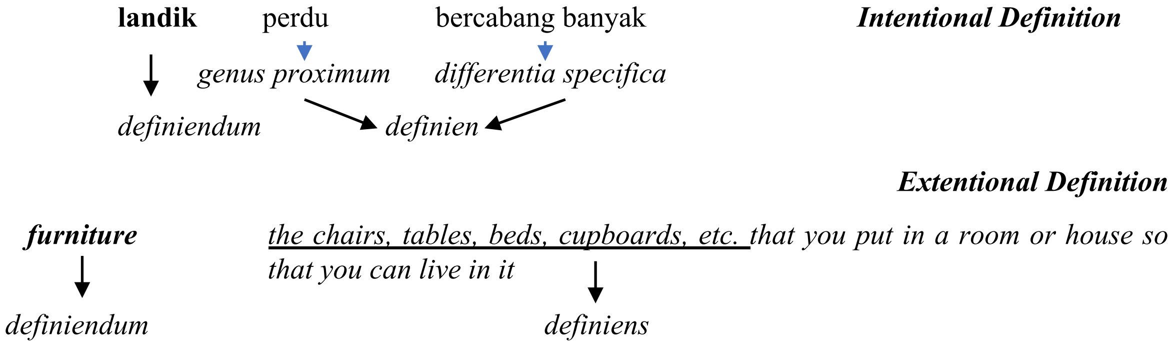
As a reference to which users frequently consult, dictionary should be compiled to meet three criteria; they are formal, functional, and content criteria. Van Sterkenburg (2003) explains that formal criterion is related to the form of the dictionary, from printed version to digital in the form of CD. Dictionary compilation should follow the suitable structure system for dictionary, in both macrostructure and microstructure. Functional criterion refers to the function of dictionary, whether it functions as a general dictionary, specialized dictionary, prescriptive dictionary, descriptive dictionary, etc. At last, the content criterion concerns linguistic information included in dictionary, such as word class, pronunciation, etymological information, etc.

In the field of lexicography, the concept of user needs in relation to the aim of dictionary compilation to meet its criteria has been developed under the function theory. According to this theory, there is a close relation between types of users, types of social situations and types of users needs that may lead them to consult dictionary (Fuertes-Olivera, 2010). In this respect, dictionary is conceived to satisfy the specific relevant needs that may arise from specific social situations of specific types of users that may lead them to dictionary consultation. In connection with dictionary user's needs for information, the information presented in general dictionary is obviously different from terms dictionary. General dictionary is intended to lay users who need to find general information. Meanwhile, terms dictionary is used by the users working in or dealing with certain expertise or field to find specific and detailed technical information. Therefore, the information contained in similar definition is different between both types of dictionary. For example, *to abut* in legal dictionary means "when two parcels of real property touch each other"; there is technical term *property* being used to define it (https://dictionary.law.com/). In contrast, *to abut* is defined by general dictionary as "to border on; to touch along the edge" (https://www.merriam-webster.com/). It contains simple and general information, there is no technical term in its *definiens*.

The main task of dictionary compiler is to catch the meaning of a word then describe it into a definition to make users understand. In dictionary, lemma represents *definiendum* (the item being defined) and definition is called *definiens* (the concepts describing *definiendum*) (Svensén, 2009). *Definiens* are arranged properly based on the function of the dictionary in order to make understandable for the users. Therefore, the definition designed for general dictionary is not similar with the terms dictionary.

In general, there are two types of definition, they are intentional and extensional. The intentional definition describes the generic conceptual relation in which the concepts are classified based on their similarities or differences. It contains hierarchic conceptual system, including superordinate, subordinate, and coordinate. The process of intentional definition involves the superordinate concept after *definiendum* or *genus proximum* (general features), followed by the information about features or distinguishing characteristics (*differentia specifica* or specific features). The more specific features included, the longer the definition described. However, an ideal definition for general dictionary only contains the most significant or core concepts and it is not encyclopedic (Svensén, 2009). Intentional definition is mostly used to define the technical terms, particularly the terms with taxonomy as those found in biology. In contrast, extensional definition is the description of *definiens* consisting the partial concepts of *definiendum* and it represents partitive conceptual relation. The examples for both definition types can be seen from Picture 1 below.

Picture 1. Types of Definition


landik
perdu
bercabang banyak
Intentional Definition
genus proximum
differentia specifica
definiendum
definien
Extentional Definition
furniture
the chairs, tables, beds, cupboards, etc. that you put in a room or house so that you can live in it
definiendum
definiens

According to Jackson (2002), there are five methods in defining *definiendum*:

1. Analytic method, the lexemes are divided based on their types or groups and their specific characteristics are described. For example, in the definition of *srigading*, *perdu* is the group to which *srigading* belongs.

   srigading perdu yang batangnya berkayu, berwarna putih kotor dan bercabang amat rapat, bunganya berbentuk malai, tabung kelopaknya berbentuk corong, sedangkan tabung mahkotanya berbentuk silinder warna oranye, buahnya bulat telur terbalik seperti jantung, tetapi pipih

2. Synthetic method, the lexeme is regarded as part of a whole unit. For example, in the definition below, *radiasi efektif malam* is regarded as a part of a group of *radiasi efektif*.

   radiasi efektif malam radiasi efektif, baik yang menghadap ke atas maupun yang menghadap ke bawah, yang terjadi saat tidak ada radiasi surya

3. Symbolization method, the definition focuses on the specific characteristics (symbol) belongs to the defined referent lexeme. For example, in the definition below, *sikudomba* is classified into sea mammal and its characteristics representing type of whale (*ikan paus*) are described.

   sikudomba mamalia laut, jenis ikan paus, giginya kecil dan bermoncong panjang, ukurannya mencapai 175—400 cm, berat 150—200 kg, hidup di perairan tropis dan subtropic

4. Rule-based method, lexeme is defined following the use convention or grammar. This method is used to define grammatical lexeme. For example:

   mengapa kata tanya untuk menanyakan sebab, alasan, atau perbuatan

5. Synonym, the definition only contains synonym or other lexeme with related meaning. For example, **hibiskus** bunga sepatu.

As mentioned before, functional theory in lexicography stresses on the approach of users as the significant factor in dictionary compilation. This study applies this approach to focus on the information needed by general dictionary users from the definition of technical terms that are usually contained in terms dictionary. It also tries to compare which definition is preferred most by the users of general dictionary KBBI V. Besides finding the preferred definition model for KBBI V, this research also makes evaluation that is significant for dictionary development and improvement.

The literature review shows that there are some previous studies about dictionary use and definition. One of them similar to the topic of this research was conducted by Amilia (2018) with *Pragmatic Aspects of Definition in Technical Terms Dictionary*. In her article, Amilia described the pragmatic aspects in defining technical words in term dictionary, including conceptual and formal aspects. She found that the definition concept of technical term only consists of a single *definien* that has the features of context, participant, norm, genre, and evident context. Despite of the discussion about how technical terms are defined in dictionary, both focus studies are quite different. Amilia only focuses on the pragmatic aspects in defining technical terms in terms dictionary. In contrast, this study aims to trawl information about the definition model of technical terms preferred by general dictionary users, particularly for the technical terms like plant and animal names. This preference refers to the target users of KBBI as a general dictionary. It means that the users are considered as lay or common users who consult KBBI because they need general information about technical terms (plant and animal names lemmas). The need of KBBI users consulting dictionary for plant and animal names is certainly different with the need of a zoologist consulting dictionary of zoology terms for the same lemmas. In addition, this study also tries to see the users' satisfaction of KBBI's definitions in general, concerning the quality, the quantity, and the presentation of information in KBBI V.

2 Method

This research is conducted by using descriptive qualitative method and functional approach in lexicography. They are applied to reveal the user lexicographical needs that lead their preference to the definition model of special lemmas in KBBI V. Those lemmas are 30 names of plants and animals whose conceptual meanings involve specialized classifications in botany and zoology. The users' needs are defined as searching necessary information from definition so that makes the users understand the concept of the lemma. This study is conducted by collecting the users' opinions through questionnaires delivered to 90 people in Bandung, West Java. The respondents are divided into two groups, academic and mass media groups. Academic group is consisted of 32 teachers, students, lecturers, and university students. Meanwhile, mass media group involves 58 journalists and editors working for publishers and printing companies.

The questionnaire used in this research is consisted of three parts. Firstly, the information for defining 30 lemmas of plant and animal names is extracted from various sources and is categorized into four categories, they are:

1. Taxonomy/scientific name, includes taxonomy rank and scientific name.
2. Characteristics, include physical features or characteristics of the referent.
3. Habitat, refers to place of living.
4. Additional information, refers to any information that is not related to the three categories.

For each lemma, the information presented is arranged in such a way so that the respondents think the information is arranged randomly. They are asked to rank the significancy of each information category on a scale of 1 to 4 (1 refers to the most significant information). The lemmas include 15 names of animals and 15 names of plants. Table 1 shows list of the lemmas.

Table 1. List of Plant and Animal Names Lemmas

No.	Plants	Animals
1.	arbei	agal
2.	bidara	alu-alu
3.	cabai	belibis
4.	dahlia	haruan
5.	enau	itik
6.	jagung	kepah
7.	kedelai	lebah
8.	keladi	pacet
9.	mangga	simpanse
10.	nanas	terubuk
11.	putri malu	uir-uir
12.	rasamala	ungkau
13.	setawar	walabi
14.	turi	wawa
15.	wijayakusuma	zebra

Secondly, for each lemma, the respondents are given four definitions to choose. For the purpose of the research, the definitions are taken from two types of dictionaries: terms dictionaries and general dictionaries. The general dictionaries used in this study are the online KBBI V and *Kamus Umum Bahasa Indonesia* (1994) or KUBI while the terms dictionaries are *Kamus Pertanian Umum* (2013) and *Kamus Istilah Dunia Peternakan* (2019). One model is arranged as a comparison, being developed from Atkins and Rundell's template entry for animal lexical set (2008). The proposed model is a full definition in which the information as *definiens* are arranged in sequent order based on the categories previously defined. For example, the information to define *corn* can be categorized as follows: tall plant whose Latin name is *Zea mays* (scientific name category); having solid stem and bearing the grain, seeds, or kernels on large ears (characteristics category); can grow in light, medium, and heavy soils (habitat); being used as food and livestock (additional information category). When these are arranged, the definition can be as follows:

corn *n* tall plant (*Zea mays*) having solid stem and bearing the grain, seeds, or kernels on large ears; can grow in light, medium, and heavy soils; usually used as food and livestock

Thirdly, the respondents are asked about their general opinion on the definitions in KBBI V. The questions are about the quantity of the information, the quality of the information, and how the information presented in dictionary.

The data collected from the questionnaire is arranged into tables on Microsoft Excel worksheet then is classified and analyzed. The researcher sorts the necessary data that can be used in this research. The irrelevant answers from the respondents are set aside then the data is recompiled systematically into descriptive text, graph, chart, and table.

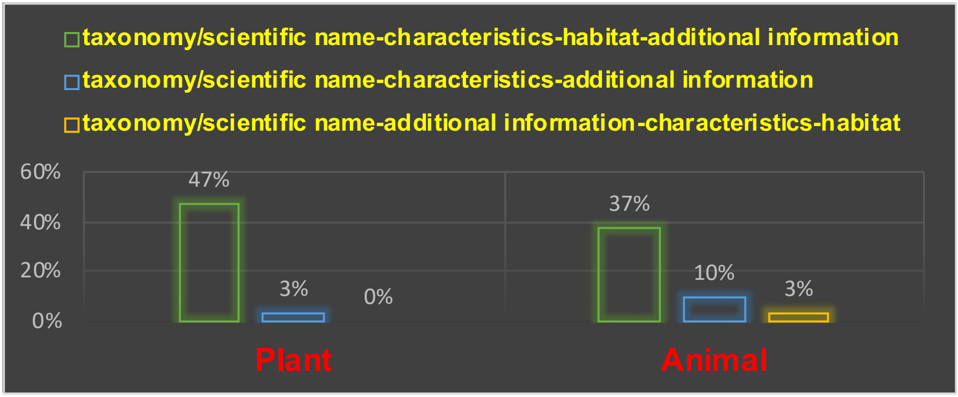
3 Result and Discussion

At the first part of questionnaire, the respondents are asked to rank four categories of information that is relevant and significant for defining lemmas of plant and animal names. The questionnaire result shows that there are three patterns mostly chosen by the respondents for the forming of relevant information contained into definition. The preferred formations are ranked in sequent order, they are 1) taxonomy/scientific name-characteristics-habitat-additional information, 2) taxonomy/scientific name-characteristics- additional information, 3) taxonomy/scientific name-additional information-characteristics-habitat.

The questionnaire result reveals the fact that for the plant names lemmas, most respondents prefer the formation of relevant dan significant information of "taxonomy/scientific name-characteristics-habitat-additional information" (47%) to the pattern of "taxonomy/scientific name-characteristics-additional information" (3%). However, no respondents choose the pattern of "taxonomy/scientific name-additional information-characteristics-habitat".

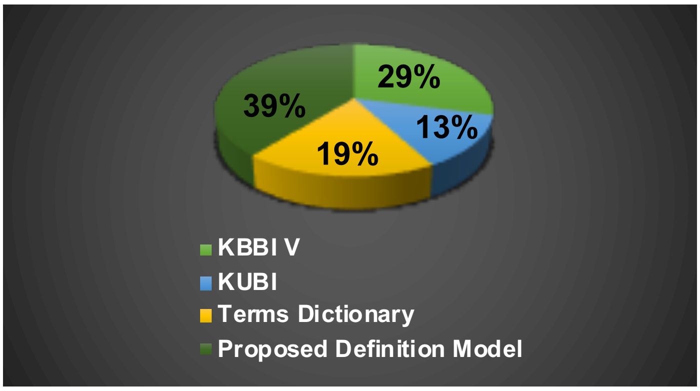
For animal names lemmas, the definition that is mostly preferred is the pattern of "taxonomy/scientific name-characteristics-habitat-additional information" (37%), followed by the pattern of "taxonomy/scientific name-characteristics-additional information" (10%) and the pattern of "taxonomy/scientific name-additional information-characteristics-habitat". Graph 1 below shows the data.

Graph 1. The preferred formation pattern of information in definition



Next, the second part of the questionnaire asks respondents about their preferences for the definition model to define lemmas of plant and animal names in KBBI. Graph 2 shows the result of the questionnaire.

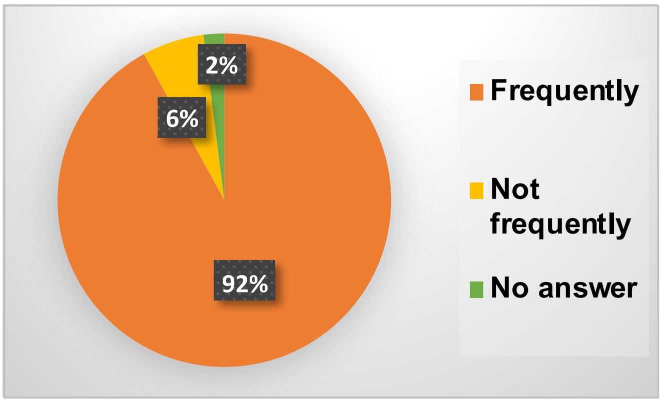
Graph 2. Preferred Definition Model



We can see from the graph above that the definition model to define lemmas of plant and animal names is the model proposed by in this study in which the information in *definiens* are arranged consistently with the formation pattern of "taxonomy/scientific name-characteristics-habitat-additional information" (39%). The KBBI's definition is the second preferred model chosen by the respondents (29%), followed by the definitions of terms dictionaries (19%) and KUBI (13%).

The third part of the questionnaire asks the respondents' general opinion on the definitions in KBBI V. The first question is about how frequent they find the information they need in KBBI V. Graph 3 below shows the result.

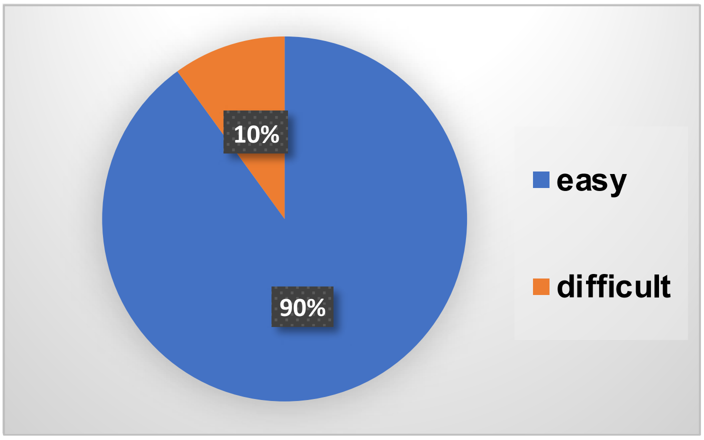
Graph 3. frequency of information found in KBBI V



From the graph we can see that 92% respondents say that they frequently find the information they need in KBBI V where 6% say otherwise. Only 2% respondents do not answer this question. It is obvious that in general, KBBI V has fulfilled the need of its users for information. It also proves that KBBI V meets the evaluation criteria related to the quantity of information.

Next, the respondents are asked whether they understand the definitions given by KBBI V. The result is presented in Graph 4 below.

Graph 4. the easiness to understand the definitions in KBBI V



The graph reveals the fact that 90% respondents easily understand the definitions given in KBBI V, whereas 10 % think otherwise. From the data, it is concluded that the definitions in KBBI V have been designed in such way to make it comprehensive and meet the clearness criteria. Nevertheless, the respondents who answer "easy" add that the definitions in KBBI V still need some improvements, especially the circular definitions. Further, they argue that the number of examples for context use should be increased to help them using the lemma in sentence.

The last question in this questionnaire is about the accuracy of the definitions in KBBI V. 93% respondents believe that the definitions in KBBI V are accurate. 7% respondents think not all definitions are accurate because they still find some definitions in which the information presented needs crosscheck with relevant experts. Despite of critics, it can be said that the definitions in KBBI V meet the evaluation criteria of accuracy.

4 Conclusion

Based on the result of the questionnaire, the researcher comes to three main conclusions. First, respondents mostly choose the pattern of concept formation as "taxonomy/scientific name-characteristics-habitat-additional information" to define both the lemmas of plant and animal names. This pattern has been applied consistently by the researcher to all definitions used in this research.

Second, the respondents consider that the information category of taxonomy/scientific name and characteristics are the most significant in a definition. This can be seen when they also choose the definition in KBBI V in spite of the irregularity of its pattern. It happens since both KBBI V and the proposal model contain the information needed most by the users.

Third, most of the respondents assume that KBBI V meets the criteria of evaluation in terms of the quality of information, the quantity of information, and how the information presented in dictionary. This proves that KBBI V satisfies the needs of its user although there's still improvement to do to make it better and fulfill the users' needs.

5 References

Amilia, Fitri. (2018). Pragmatic aspects of definition in technical terms dictionary. *Budapest International Research and Critics Institute*, 1 (3). http://bircu-journal.com/index.php/birci/article/view/51

Atkins, B.T. (Ed.). (1998). *Using dictionaries: Studies of dictionary use by language learners dan translators*. Germany: De Gruyter.

Badudu, J.S. dan Zain, Sutan Mohammad. (1994). *Kamus umum bahasa Indonesia*. Jakarta: PT Intergrafika.

Bergenholtz, Henning. (2003). User-oriented understanding of descriptive, proscriptive and prescriptive lexicography. *Lexikos*, 13, 68. https://doi.org/10.5788/13-0-722

Bielinska, Monika dan Schierholz, Stefan J. (Eds.). (2017). *Dictionary criticism*. Germany: De Gruyter.

Bozkurt, Ferdi. (2011). Technical word boundary for general purpose dictionary: A grounded theory. *Proceedings of the 11th International Conference of the Asian Association for Lexicography*, Guangzhou, 211. https://www.researchgate.net/publication/325385061_Technical_Words_Boundary_for_General_Purpose_Dictionaries_A_Grounded_Theory

Budiarsa, Komang. (2019). *Kamus istilah dunia peternakan*. Sidoarjo: Zifatama Jawara.

Creswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. London: Sage Publications.

Jackson, Howard. (2002). *Lexicography: An introduction*. London: Routledge.

Lehmann, Christian. (4 Maret 2020). Functions and types of dictionary. https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html?https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/functions_and_kinds.html

Lew, Robert. (2015). Opportunities and limitations of user Studies. *Online Publizierte Arbeiten zur Linguistik*, 2. https://ids-pub.bszbw.de/frontdoor/deliver/index/docId/3772/file/Research_into_dictionary_use-2015.pdf

MacMillan, J.B. (1949). Five college dictionaries. *College English* 4, 214—21.

Miles, Mathew B. dan Hubermn, A. Michael. (1984). *Qualitative data analysis: An expanded sourcebook*. London: Sage Publications, Inc.

Moleong, Lexy J. (2007). *Metodologi penelitian kualitatif*. Bandung: Remaja Rosda Karya.

Muller-Spitzer, Carolin. (2008). Research on dictionary use and the development of user-adapted views. *Text resources and lexical knowledge: Selected papers from the 9th Conference on Natural Language Processing Konvens 2008*, 223—228.

Olivera, Pedro A. Fuertes. (Ed.) (2010). *Specialized dictionary for learners*. Germany: De Gruyter.

Sugiyono. (2008). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.

Suryana. (2007). *Tahap-tahap penelitian kualitatif mata kuliah analisis data kualitatif*. Bandung: Universitas Pendidikan Indonesia.

Sutopo, H.B. (2006). *Metode penelitian kualitatif*. Surakarta: UNS Press.

Svensén, Bo. (2009). *A handbook of Lexicography: The theory and practice of dictionary-making*. United Kingdom: Cambridge University Press.

Tarp, Sven. (2009). Reflections on lexicographical user research. *Lexikos*, 19. https://doi.org/10.5788/19-0-440.

Tim Penyusun Kamus. (2013). *Kamus pertanian umum*. Jakarta: Penebar Swadaya Grup.

Van Sterkenburg, Piet. (Ed.) (2003). *A Practical guide to lexicography.* Leiden: John Benjamins Publishing Company.

Welker, Herbert Andreas. (2010). *Dictionary use: A general survey of empirical studies*. Brasil: Author's Edition.

Yong, Heming dan Jing Peng. 2007. *Bilingual lexicography from a communicative perspective*. Amsterdam/ Philadelphia: John Benjamins.

OBSOLETE, ARCHAIC, AND CLASSIC WORDS IN INDONESIAN DICTIONARY: A PRELIMINARY INVESTIGATION

Dewi Puspita, Kamal Yusuf
National Agency for Language Development and Cultivation, Indonesia;
UIN Sunan Ampel Surabaya, Indonesia
dewi.puspita@kemdikbud.go.id; kamalyusuf@uinsby.ac.id

Abstract

The global era has led to a fairly rapid change in language. Many words have become obsolete. There are also many words whose meaning becomes irrelevant nowadays. Unfortunately, in Indonesian dictionary, especially in Comprehensive Indonesian Dictionary (KBBI), there is no label for obsolete words. There are only archaic label to mark all outdated words and classic label to mark classical words. Another labeling problem in the KBBI is that there are no clear guidelines or criteria to determine when a word is considered archaic, obsolete, or classic. The absence of clear criteria causes some entries that have been labeled archaic in the KBBI to seem obsolete and sometimes classic words get confused with the archaic. The aim of this paper is to investigate how to categorize archaic, obsolete, and classical words in KBBI. This research was conducted by comparing several forms and entry criteria labeled archaic, obsolete, and classical in several dictionaries, especially dictionaries of foreign languages whose lexicographic tradition had been established. Each dictionary has its own criteria for classifying a word as archaic, obsolete, or classical, and we can learn from them. From the results of our study, we suggest that, for now, the best and easiest way to categorize words according to their labels is to check the corpus database.

Keywords: obsolete, archaic, classical, Indonesian dictionary

1. Introduction

People are more likely to care about new words. Many new words have been coined to name new concepts. Every year there are about a dozen new popular foreign (especially English) terms or words that come to Indonesian through the media and most of them are adapted into Indonesian. If those adapted new terms or words last long enough or become very popular in Indonesian society, there is a big chance that those word will be included in the Indonesian dictionary. Those people forget or may not be aware that in Indonesian there are old words that may already contain the same concept. Old words should also receive the same attention.

Recently, we found pictures (as can be seen in appendix) on social media containing Indonesian words that have not been used for a long time and are no longer recognized by at least two generations of Indonesian speakers. Those words, to mention some, are:

acaram ‘ring’	*angga* ‘deer antlers’
barid ‘messenger’	*cegar* ‘cascade’
daayah ‘propaganda’	*darpana* ‘hand mirror’
galiung ‘big sailing ship’	*khat* ‘(beautiful) handwriting’
intikad ‘critique’	*jambar* ‘dish’
khauf ‘fear; worry’	*kutubkhanah* ‘library’
nazim ‘writer; poet’	*setanggi* ‘incense’
syabas ‘a shout of approval’	*tanju* ‘wall lamp’
teja ‘yellowish red light at sunset’	*wau* ‘paper kite’

All the words mentioned above as depicted in the picture can be also found in the Comprehensive Indonesian Dictionary or Kamus Besar Bahasa Indonesia (KBBI), starting from the first edition to the last one (5th edition), which means that these words have been used in certain time in Indonesian. KBBI is a general dictionary and also a historical dictionary. It records all linguistic facts that have been and are currently living in the Indonesian speaking community. That is why even old words that were no longer used like the words above, are still listed in the dictionary.

There are five language variant labels in the current edition (5th) of KBBI. They are:

1. archaic (*ark*), '*arkais*', is given to words that are no longer commonly used,
2. classic (*kl*), '*klasik*', to mark words that are listed in classical Malay literature,
3. respectful (*hor*), '*hormat*', to marks words that are used in formal situation,
4. colloquial (*cak*), '*cakapan*', is used to mark words that are usually used in an informal situation,
5. rude (*kas*), '*kasar*', refers to words that are considered rude or impolite to use in normal conversation.

Some of the aforementioned words are labeled archaic such as *darpana* and *intikad*. Some are labeled as classic such as: *acaram*, *tanju* and *galiung*, but some others had no labels. In the history of Indonesian dictionaries, after the official language in Indonesia is called Indonesian, the monolingual dictionaries that are often used as valid references are General Dictionary of Indonesian *'Kamus Umum Bahasa Indonesia*' (Poerwadarminta, 1952), General Dictionary of Indonesian '*Kamus Umum Bahasa Indonesia*' (Badudu and Zain, 1983), Dictionary of Indonesian '*Kamus Bahasa Indonesia*' (Adiwimarta, 1994), and all editions of Comprehensive Indonesian Dictionary or KBBI (the first edition 1988, the second edition 1991, the third edition 2000, the fourth edition 2008, and the fifth edition 2016). In the reference dictionaries published before 2000, there are only three labels of language variety, namely *cak*, *hor*, and *kas*. Label for archaic and classic words just started appearing in the third edition of the KBBI. The following table presents data on the number of words labeled obsolete, archaic, and classic in all edition of KBBI.

Table 1. Numbers of entries labeled obsolete, archaic, and classic in KBBI

KBBI edition	**Label**		
	obsolete	**Archaic**	**classic**
1st edition (1988)	0	0	0
2nd edition (1991)	0	0	0
3rd edition (2000)	0	1227	991
4th edition (2008)	0	1253	977
5th edition (2016)	0	1258	1025

Although it has been explained that there are no archaic and classic labels in the first and the second edition of KBBI, In Table 1, all data are shown, including zero (0) data, in order to see a contrast of data changes. The number of entries labeled archaic jumped from 0 in the first and the second editions to 1,227 in the third edition, while the additions were insignificant in the fourth and the fifth editions. Likewise with classic labels. The number of entries labeled classic increased intensely from 0 in the first and the second editions to 991 in the third edition. However, this figure was reduced by 14 in the fourth edition to 977. In the fifth edition the number of entries labeled classic increases again to 1025.

Unfortunately, the preface of all edition of KBBI does not mention what criteria are used when labeling entries based on these various language varieties. KBBI does not explain when a word can be categorized as archaic. Somehow, obsolete label is not found in KBBI. All words that are no longer used or outdated are considered archaic. In fact, almost all of the words labeled classic have not been used for a very long time. Many words with classic labels should also be archaic. There is no clear definition for each label which causes confusion for language users. In addition, there are no clear categories for labeling. Due to this obscurity, many entries are mislabeled. There is no explanation of why *acaram* and *darpana* are labelled archaic while *khat*, *khauf*, and *teja* are not. Then what about words whose objects or concepts no longer exist? Are they also archaic?

The decision to determine the labeling of words often rests at the discretion of each dictionary compiler and the purpose of the dictionary. So it is with KBBI. However, the labeling criteria should have been explained in the preface or in the instructions for use so that the dictionary user knows the mechanism. Dictionary users can also provide suggestions for improvements and benefits of the dictionary. In this article we tried to investigate how to categories the three language variety labels in KBBI, namely obsolete, archaic, and classic, so that their differences can be clearly seen. The categorization that has been obtained will be proposed to KBBI compilers so that the classification and labeling of words in KBBI become clear and no longer overlaps.

From the aforementioned description, this paper is aimed to shed light out the obsolete, archaic, and classic words in Indonesian dictionary (KBBI). Specifically, this study investigated how to categorize archaic, obsolete, and classical words in KBBI. This is important to write this article with the intention of bringing back old words in the dictionary to the public's concern.

2. Literature Review

Just as the old vocabulary that has rarely been noticed, there are not many literature and theories regarding labeling obsolete and archaic words. Muggelstone (2000) said that labeling is a problematic area for lexicographers because its consistency and consensus are difficult to achieve. Labeling is also a subjective judgement of lexicographers (Brewer, 2015). Regarding this subjectivity, the Oxford English Dictionary (OED) chief editor in OED general explanations stated that:

> […] "And the death of a word is not an event of which the date can be readily determined. It is a vanishing process, extending over a lengthened period, of which contemporaries never see the end. […] It is only when no one else is left to whom its use is still possible, that the word is wholly dead. Hence, there are many words of which it is doubtful whether they are still to be considered as part of the living language; they are alive to some speakers, and dead to others."

Based on the statement above, not many dictionaries, even of major languages that their lexicographic tradition have been long established. They include archaic or obsolete words as their entries. Maxwell (2006) stated that "The OED is the only dictionary which retains entries for every single word that has ever existed in the English language. Most other dictionaries routinely exclude words that are outdated or obsolete." Maxwell's statement is not entirely correct. Besides OED, there are several other English dictionaries that still retain obsolete words as their entries. Merriam Webster's Dictionaries, unabridged and collegiate editions, still include obsolete and archaic words. In Prentis' article (2017), managing editor of the unabridged Collins English Dictionary, Marry O'Neil said "We rarely take words out of our dictionaries. This is especially true of our larger dictionaries. If we find that a word has fallen out of general use, or is not used as much as it was before, we usually label such words as 'obsolete,' 'archaic,' or 'old-fashioned' rather than deleting them entirely."

The presentation of labels in obsolete and archaic entries in various dictionaries provide us with an understanding of good techniques for labeling and presenting entries with those labels in a dictionary.

3. Method

This research was conducted qualitatively with the literature study method. Theories of dictionary labeling is studied. The labeling criteria used by several dictionaries are also studied and compared to obtain a comprehensive understanding. Theories and labeling of other dictionaries were then compared to the labeling in the Comprehensive Indonesian Dictionary (KBBI). From the literature study and comparisons we can determine what criteria a word must meet in order to get an obsolete, archaic, or classical label.

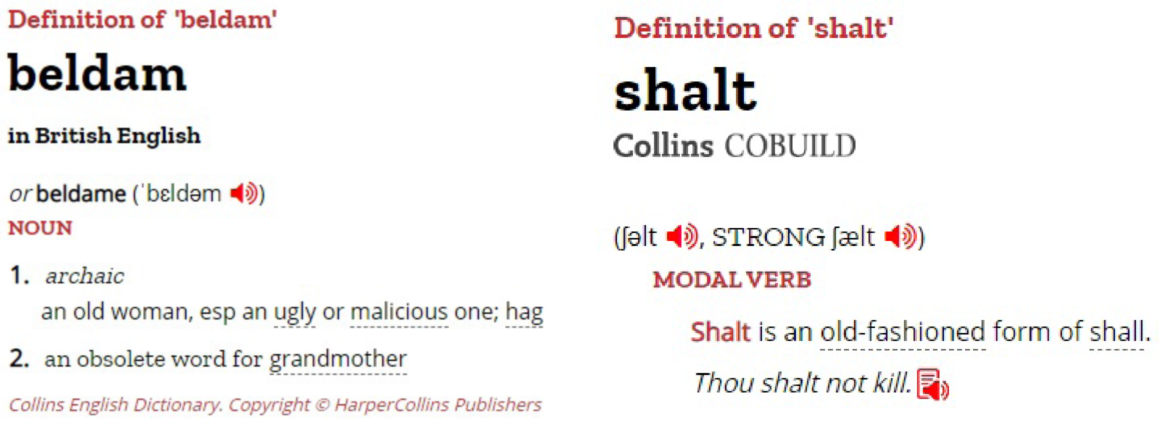
Data which included the archaic and classical words in this study were taken from the KBBI entries. The data is tested through several corpora to see the degree of archaic or classical or the accuracy of the labeling of the words. The corpora used is a variety of Indonesian language corpora available, including the web as corpus. From the test results it will be seen whether the label that has been given is correct, or needs to be changed. From the results, it can also be seen whether the obsolete label needs to be added to the KBBI or not.

4. Result and Discussion

As follows, we present the results of the study and discuss the following findings.

Each Dictionary Gives Different Labels to Outdated Words

OED labels words based on their status as follows: *obs*. (obsolete), *arch*. (archaic or obsolescent), *colloq*. (colloquial), *dial*. (dialect). There is also a marker for rare word with a $^{-1}$ or $^{-0}$ sign which indicates that only 1 or 0 actual instances of the use of the word is known to the lexicographers. words that are coined for one occasion only (nonce word) are marked with the label *nonce wd*. Obsolete words as well as obsolete meanings are marked with a dagger in front of them to distinguish them from words that are still actively used. Collins Online Dictionary does not create special labels with abbreviations for obsolete, archaic, and old-fashion words, but includes the label in the definition. Examples can be seen in the two images below.



Definition of 'beldam'

beldam

in British English

or **beldame** (ˈbɛldəm)

NOUN

1. *archaic*
   an old woman, esp an ugly or malicious one; hag
2. an obsolete word for grandmother

Collins English Dictionary. Copyright © HarperCollins Publishers

Definition of 'shalt'

shalt

Collins COBUILD

(ʃəlt, STRONG ʃælt)

MODAL VERB

Shalt is an old-fashioned form of shall.

Thou shalt not kill.

Figure 1 word definition labeling in Collins Dictionary

Malaysian Comprehensive Dictionary (Kamus Dewan) in its fourth edition (2015) has only one label for its dictionary entires, namely archaic (ark). This label is not found in previous editions. In its third edition (1998, xli), words that met the archaic criteria were marked with dagger symbol (†). The symbol is no longer appear now.

Classic Label Is Only Found in KBBI

Apart from KBBI there is no other dictionary that label its entry as classical, including Kamus Dewan Malaysia. English as well as other languages that have been around for a long time, has loads of classical literature. However, the vocabulary found in those classical literature is not specifically labeled classic when it is included in the dictionary entry. Likewise with the Malay language, Indonesian and Malaysian share most of the same classical Malay literature. However, the Malaysian dictionary does not mark the vocabulary of classical literature as a classic word, but as an archaic.

The consideration of the compilers of the third edition of KBBI in labeling vocabulary derived from old Malay literature as a classic word may be because Indonesian vocabulary is now increasingly different from Malay vocabulary. The problem is, the decision to label the word as classic adds confusion to the categorization. Most of the classic words are archaic and some are already obsolete. In addition, most of the classic labelled words are also general, not literary-specific.

Labeling Criteria

All of the dictionaries used in this study have their own considerations and criteria when deciding to label an entry. The criteria used by some dictionaries for obsolete and archaic labeling are as follows.

The determination of obsolete labels in the OED refers to its database of 2.5 billion words. In OED, if no citation evidence is found for an item dating from 1930 or later, the item is labeled obsolete. Words that are no longer used today, but which are still useful to historians or historical novelists, would be labeled *Now hist*. This is explained by Gilliver (2016) in his book *The Making of the OED*, as quoted in https://qz.com/1061782/the-complex-process-that-dictionaries-use-to-decide-which- words-are-obsolete/

There are three types of status labels used in Merriam Webster's Online Dictionary, namely temporal, regional, and stylistic. In temporal, there are two labels: obsolete and archaic. The obsolete label is given to a word that has no evidence of its use since 1755. The archaic label is given to a word or sense that once common in use but today it is found only sporadically or in special contexts.

The Collins Online Dictionary defines archaism by using a large database analysis of written and spoken language gathered from various sources to assess whether or not a word is still widely used. The database is continuously updated until now there are about 4.5 billion words in it.

The fourth edition of Kamus Dewan Malaysia (2015) put an archaic label for the word which has the following condition:

- rarely used,
- its use is limited to one area or environment,
- ancient or dead,
- its accuracy is questionable (it can be caused by misreading, miswriting, mishearing, etc.).

Since ancient and dead words fall into the criteria for words labeled as archaic, there is no need for obsolete label in this dictionary.

Obsolete Label in Indonesian Comprehensive Dictionary Should Exist

Indonesian Comprehensive Dictionary (KBBI) considers all outdated words to be archaic, in fact archaic words will become obsolete and dying out. There must be a distinction between words that are no longer used and words that are dead.

Determining Obsolete, Archaic, and Classic Label Criteria through Corpus Checking

A corpus can be used to see the frequency at which words occur. This feature can be used to check whether a word does not appear in the corpus; has appeared in historical corpus but does not appear again in recent corpus; or only appear in certain texts. In this study we suggest that corpus can be used to check whether a word can be called obsolete, archaic, or classic.

To check for obsolete words or words with obsolete meaning and archaic, we can use historical corpus. Obsolete words or words with obsolete meaning should no longer appear in the current corpus. Indonesian does not yet have an adequate corpus database for this examination. However, we can combine the use of several existing corpora, including the web as corpus. We can customize the search period in the google search engine to see in which period a word is used or no longer used.

For classic words of Indonesian, there is a classical Malay corpus from Malay Concordance Project (MCP). MCP comprises of 165 classical Malay texts from over 150 sources of pre-modern Malay written text. If it is true that the words that are labeled *kl* (classic) come from classical Malay manuscripts, then they should appear in the corpus. If a classic labelled word cannot be found in the corpus of classical Malay texts, then the validity of the label must be questioned. On the other hand, if there is an obsolete unlabeled word found in the corpus of classical Malay, the word can be labeled classic.

The search for the word *tanju* ‘wall lamp’ (a classic word labelled *kl* in KBBI) in MCP returns 8 occurrences with 7 occurrences of the same meaning. Meanwhile, the word *galiung* ‘a large sailing ship’ which also labelled *kl* in KBBI cannot be found in MCP. The checking results on the web as corpus (via google and Bing search engines) for the word *galiung* shows that the word appears only on Wikipedia, pages containing KBBI definitions, and translated text from English that use *galiung* as the equivalent of galleons. *Galiung* or galleons in English is indeed a sailboat from centuries ago, but the word is not found in classical Malay manuscripts. Therefore, the *kl* label for this word is not right. One more word labeled *kl* in KBBI in the attached image is *acaram*. In the MCP corpus, the word *acaram* is only found in the law of Melaka in the context of a marriage dowry. No other classical texts contain this word. Likewise on the web as corpus, the word *acaram* can only be found in the KBBI entry. This suggests that this word should be more accurately labeled obsolete. Corpus checking is an effective way of knowing this.

Archaic, or classical words may still be found in the corpus with minimal frequency of use. This is possible because some people like poets sometimes use old words to express their thoughts in their poetry. There are also people who deliberately bring these old words back to life, such as those who made the pictures in the appendix of this paper. Sometimes, old or classic words just sound more beautiful than the words we are used to.

Thus, we propose that a word or meaning can be labeled obsolete if the word or meaning has completely disappeared and is no longer found in the corpus database of at least the 21st century. Meanwhile, a word can be labeled archaic if it is outdated but can still be found in the current corpus data, such as in prose or poetry, in very small quantities. Furthermore, classic labels can be used to mark words that do contain classical values, not just to label words that appear in classical literature. This requires a very large Indonesian language database.

What to Do If the Proper Labeling Cannot Be Decided

If it is difficult to determine the right label for a word, it is best to add an explanation to the definition, as an example of defining the word galleon in the OED below: ‘A sailing ship in use (especially by Spain) from the 15th to the 18th centuries, originally as a warship, later for trade. Galleons were typically square-rigged and had three or more decks and masts’ https://www.lexico.com/definition/galleon. In that definition, it is explained that the galleon is from the 15th to the 18th century. The explanation of the period of use of the word can also be presented in graphic form as is done by the Collins English Dictionary in https://www.collinsdic tionary.com/. The Collins has a feature that can trace a word’s usage back over 10, 50, 100 or 500 years to discover how its popularity has changed over time. In the graph below the period of use of the word galleon in Collins’ database is presented. From the graph we can see that the word *galleon* was widely used in the 1700s, after which the frequency of its use decreased.

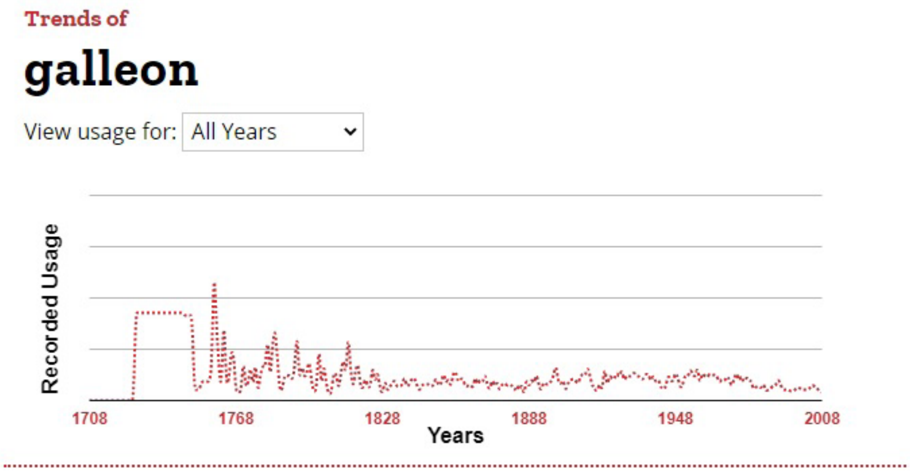


Figure 2 The trend of word Galleon in use

Source: https://www.collinsdictionary.com/dictionary/english/galleon

Explanation in definition and / or through graphs like the one above makes it easier for dictionary users to understand the concept of the word than just to read a classical label with the short definition of 'a large sailing ship'. As an online dictionary, Indonesian Comprehensive Dictionary can make similar or more informative features according to its database.

5. Conclusion

The main thing that can be concluded from this study is that determining labels for old words is not easy, it really depends on the subjectivity of the dictionary compiler. However, general criteria can be established so some guidelines that the next generation of dictionary compilers can follow. In addition, even if there was a labeling error, it would not have strayed too far from what it should have been.

The current article is a preliminary study. There is still much to be studied and explored on this topic. The research method can be added not only in the form of literature study but also combined with interviews with several generations of language speakers regarding how far they are familiar with vocabulary that is considered archaic or obsolete. At least, from the results of this study, KBBI compilers can start to reconsider the existing labeling in the dictionary, consider to add obsolete labels, review the classic labels, and double-check whether the labeled words are already correct.

References

Adiwimarta, S. S., et al. 1983. *Kamus Bahasa Indonesia*. Jakarta. Pusat Pembinaan dan Pengembangan Bahasa.

Badudu, J. S. and Zain, S. M. 1994. *Kamus Umum Bahasa Indonesia.* Jakarta: Pustaka Sinar Harapan

Brewer, C. (2016). Labelling and Metalanguage. In Durkin, P. (Ed.). (2016). *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Card, W., McDavid Jr, R. I., & McDavid, V. (1984). Dimensions of Usage and Dictionary Labeling. *Journal of English Linguistics*, 17(1), 57-74.

Collins Online Dictionary . 2021. Accessed from https://www.collinsdictionary.com

Dike, E. (1935). Obsolete English Words: Some Recent Views. *The Journal of English and Germanic Philology, 34*(3), 351--365.

Durkin, P. (Ed.). (2016). *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press. Galleons. 2021. Accessed from https://www.lexico.com/definition/galleon

Gilliver, P. 2016. *The Making of the Oxford English Dictionary*. Oxford: Oxford University Press.

Kamus Besar Bahasa Indonesia. 1988. 1st Edition. Jakarta: Pusat Bahasa
Kamus Besar Bahasa Indonesia. 1991. 2nd Edition. Jakarta: Pusat Bahasa
Kamus Besar Bahasa Indonesia. 2000. 3rd Edition. Jakarta: Pusat Bahasa
Kamus Besar Bahasa Indonesia. 2008. 4th Edition. Jakarta: Pusat Bahasa
Kamus Besar Bahasa Indonesia. 2016. 5th Edition. Jakarta: Pusat Bahasa

Kamus Dewan. 2015. 4th Edition. Kuala Lumpur. Dewan Bahasa dan Pustaka Kamus Dewan. 1998. 3rd Edition. Kuala Lumpur: Dewan Bahasa dan Pustaka

Maxwell. K. 2006. *The Monthly Webzine of the Macmillan English Dictionaries. Issue 38.* http://macmillandictionaries.com/MED-Magazine/May2006/38-New-Word.htm

Mugglestone, L. (2000). Labels Revisited: Objectivity and the OED. *Dictionaries: Journal of the Dictionary Society of North America*, 21, 22-36.

Poerwadarminta, W. J. S. 1952. *Kamus Umum Bahasa Indonesia*. Jakarta: Balai Pustaka.

Prentis, N. (2017). The Complex Process that Dictionaries Use to Decide which Words Are Obsolete. Accessed from https://qz.com/1061782/the-complex-process-that- dictionaries-use-to-decide-which-words-are-obsolete/

Appendix Source:

https://www.facebook.com/photo?fbid=3493834833977869&set=pb.100000541617487.- 2207520000

Acaram | اچارم
Angga
Cegar | چگر
Daayah | دعاية
Darpana | درفان
Galiung
Khat
Intikad
Jambar | جمبر
Khauf | خوف

THE USE OF VERB VALENCY PATTERNS IN THE INDONESIAN MONOLINGUAL LEARNER'S DICTIONARY

Dora Amalia
National Agency for Language Development and Cultivation, Indonesia
dora.amalia@kemdikbud.go.id

Abstract

One of the most confusing aspects for BIPA (*Bahasa Indonesia bagi Penutur Asing* 'Indonesian for Foreign Speakers') learners is the verb. Verbs in Indonesian are very rich in affixes with their various senses respectively. This makes it difficult for BIPA students to use verbs with the right affixes and word structures. Therefore, we need practical grammar information that can help these students. Verb valency patterns are a piece of short and easy- to-understand grammar information. This study discussed how to present relevant grammatical information by looking at the mistakes shown by BIPA learners from the collected learner's corpus. It starts with identifying the error in using verbs in the BIPA learner's corpus, classifying, and analyzing them. This study aims at designing verb patterns as practical grammar information presented in a monolingual learner's dictionary. The verbs studied are transitive ones with complex affixes.

Keywords: *learner's dictionary, learner's corpus, verb valency pattern, transitive verb, verb affixes*

1 Introduction

One of the most complicated things in learning Bahasa Indonesia bagi Penutur Asing(henceforth, BIPA) or Indonesian for foreign speakers is the verbal formation and use (Madia, 2001; Widawati 2002). According to *Tata Bahasa Baku Bahasa Indonesia* (henceforth, TBBI) or Indonesian Standard Grammar, there are 14 affixes used in the verb formation. The affixes consist of 6 prefixes (*meng*-; *di*-, *ter*-, *per*-, *ber*-, and *se*-); 2 suffixes (-*kan* and -*i*); 2 confixes (*ber* -/-*an* and *ke*-/- *an*); and four infixes (-*el*-, -*er*-, -*em*-, and -*in*-). Besides the fourteen affixes, verbs in Indonesian also are formed using reduplication. The fourteen affixes undergo a unique morphophonemic process when combined with a root word. Apart from the normal process, some exceptions are not by the provisions of the formation of the verb. It is additional work for BIPA students to memorize these irregular shapes. It is necessary to find a useful way for BIPA learners to master how to use verbs correctly. Each affix contains its semantic meaning. To use it appropriately, the learner must know the context in which it locates. In addition to that, how to use the verbs in the correct context is also a problem in itself.

Below is a chart of verb affixation in Indonesian. Some affixes are used separately as a prefix or a suffix, but he joins other affixes to become a confix.

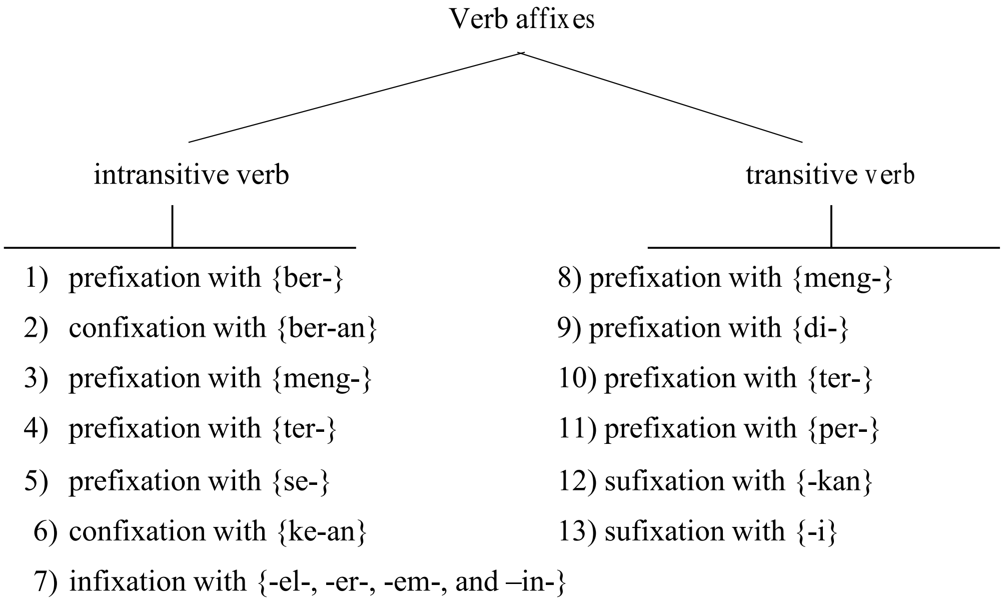

Verb affixes
intransitive verb
transitive verb
1) prefixation with {ber-}
2) confixation with {ber-an}
3) prefixation with {meng-}
4) prefixation with {ter-}
5) prefixation with {se-}
6) confixation with {ke-an}
7) infixation with {-el-, -er-, -em-, and –in-}
8) prefixation with {meng-}
9) prefixation with {di-}
10) prefixation with {ter-}
11) prefixation with {per-}
12) sufixation with {-kan}
13) sufixation with {-i}

Fig 1.1 Verb affixes

The affixation is not always a single process but multilevel or complex ones. The verb formation can involve both derivational and inflectional affixes, respectively. An example of this type of word formation is the word *memberdayakan* ‘to empower’ as follows.

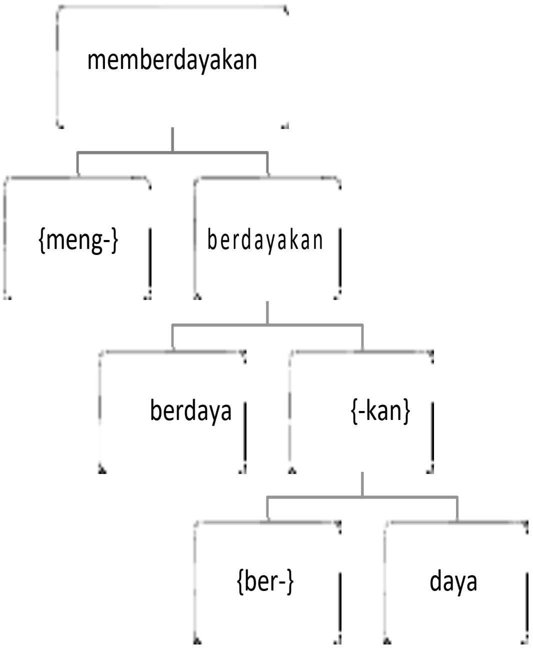


Fig. 1.2 Verb formation with a complex affixation

In the word-formation paradigm, the derivational prefix {ber-} is initially added to primary base *daya* ‘power’ and become *berdaya* ‘to have power/powerful as the secondary root. This secondary root gets the addition of the suffix {-kan} so that it produces *berdayakan* ‘empower’. The verb *berdayakan* itself becomes the tertiary root of the affixation {meng-} which results in *memberdayakan*.

As mostly perceived, the dictionary is a language learning tool that is very popular and commonly used. The functions of dictionary are receptive and productive. Regarding the use of verbs in the correct context, this productive function is the most relevant. This study will discuss how the dictionary provides syntactic information that the learners can use to produced native-like speaking or writing.

The definition is the most important part of the monolingual dictionary entry (Fontenelle, 2008) and the skill of a lexicographers is judged by how they define it. Definitions have a special status in the dictionary which have been considered as the duties of a true professional lexicographer, as Lew and Dziemianko (2006) stated "*definition ... seems to enjoy a priveleged status ... writing definitions is seen as the prototypical task of professional lexicographer*". In line with that, Rundell (2008) states that writing definitions is the biggest challenge for lexicographers, especially those who work on student dictionaries ("*… this is probably the biggest single challenge of pedagogical lexicography*").

In a productive context, the definition must contain instructions on how to use a word, in this case, a verb, in the proper context. Therefore, the definition, in this case, is a central and essential part. The development of features and innovations in LD has more or less led to the standardization of the arrangement. Rundell (2006) lists the main features that characterize LD. These features go into two groups, a receptive and a productive function. As for productive purposes, an LD is better to fulfill the following matters.

1) Syntax information is provided with the valency pattern.
2) Example sentences are provided to show the context of use and as a model for producing text.
3) Sociolinguistic features are considered.
4) Additional information such as usage notes are included.

Among these features, Rundell pays close attention to three elements that he considers the most distinguishing MLD from other dictionaries. The three elements are definition, syntactic information, and examples. Language learners need a model to imitate in producing text. Therefore, for productive needs, the example sentences are imperative in MLD. It said that the example sentences in MLD must fit the pedagogical purposes.

Regarding the productive need of BIPA learners, Amalia's study (2014) suggested the following proposal.

1) Dictionary users need definitions of traditional types and valence patterns for receptive purposes.
2) Dictionary users need the definition with valency pattern for productive purposes.
3) The information needed in the dictionary is collocations, the context of usage, and examples of typical sentences.
4) The valency pattern provides a model for making grammatical sentences.This study aims to make verb valency patterns as practical grammar information presented in a monolingual learner's dictionary.

2 Method

This research is based on a learner corpus formed from collected essay assignments of BIPA learners from various levels and native languages. There are 67 essays selected and 11 countries of origin for BIPA learners. Of the different types of errors identified, the only misuse of transitive verbs with complex affixes is discussed. Based on the identification, the errors are categorized according to their syntactic behavior.

Each category is represented by one of the most prototypical misuses. It is then corrected and modified into potential sentences. The next step is to run the semantic frame analysis so that the valency pattern of the verb is obtained.

Further analysis is conducted on a larger corpus. In this study, Sketchengine was used to collect potential meanings that had not been found in the learner's corpus. The concordance line and collocation command will show the specified words appearing together with the verb. By analyzing the collocation and the context in the concordance line, we can determine the potential senses. The results of the analysis then become the basis for compiling verb entries of MLD with practical grammatical instructions. Entry models are created for each verb.

3 Result

The results of this study is an entry model of the verb *mempersembahkan* that represent the most complex affixation. The entry model with a valency pattern is followed by other verb entries that represent a unique affix, namely {meng-/-kan} as in *membersihkan* 'to clean'; {meng-/-i} as in *mencintai* 'to love'; {memper-} as in *memperlancar* 'to expedite'; {memper- /-i} as in *memperbaiki* 'to fix'; and {member-/-kan} as in *memberlakukan* 'to apply'. The six entries are defined using all navigation tools, i.e. reduced definition, usage note, etc.

4 Analysis and Discussion

Fillmore (1976) introduced the theory of frame semantic of how to describe words. The rationale for this theory is that we can only understand the word meaning based on the semantic frame and context. The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the described word (Fillmore 1995).

Alan (2001) understands the semantic framework as a series of linguistic facts that indicate the features, attributes, and functions of a denotatum and their distinctive interactions with other things associated with them *("... is a collection of facts that specific characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it* "). For example, the concept of **cooking** usually involves a person who cooks (**Cook**), cooked food (**Food**), something to put food when it is cooked (**Container**), and a source of fire for cooking (**Heating_ Instrument**). (**Cook**), (**Food**), (**Container**), and (**Heating_Instrument**) are frame elements (henceforth, FE). All FEs represent a semantic frame (APPLY_HEAT). The words that trigger the appearance of this framework (APPLY_HEAT), such as *fry*, *cook*, *bake*, *boil*, *broil*, etc., are called LU.

In pedagogical lexicography, dictionaries should provide information on the word meaning and its use in a correct situation. According to Atkins and Rundell (2008: 147) "language learners must know how these FEs are expressed grammatically, or they cannot use the word correctly". For this reason, the definition in the learner's dictionary is more contextual. Grammatical context and collocation are the main concerns in this type of dictionary so that learners know the syntactic behavior of a word in its context. Apart from that, both users can make it a model for producing text in the target language.

Atkins and Rundell (2008) further developed the Frame Semantic theory by making valency patterns as grammatical information in dictionary entries. In their description, Atkins and Rundell describe the analytical steps carried out on the corpus to arrive at the valency pattern in the entry. The analysis as follows. *Jo asked her brother to help her.*

LU *ask* in the sentence above is included in the order (REQUEST). The actor who asks is (Jo) and the person who is asked for help is (her brother) to do what is asked (to help her). Thus, it can be determined FEs in order are as follows.

(***Speaker***) *Jo*

(***Addressee***) *her brother*

(***Message***) *to help her*

FE analysis in the semantic framework has similarities with the analysis of the role of semantics in Indonesian sentences. In this semantic role, sentence elements can be divided into (**Actor**), (**Experience**), (**Action**), (**Target**), (**Beneficiary**), (**Place**), (**Tool**), and so on. The semantic role is one of the three analyzes in describing sentences besides the analysis of the grammatical and categorical functions

The next step is to annotate based on the type of phrase and grammatical function. This process is the same as describing functions and syntactic categories in sentences. In this process, each sentence element is given information according to its function and category. The complete annotation as follows.

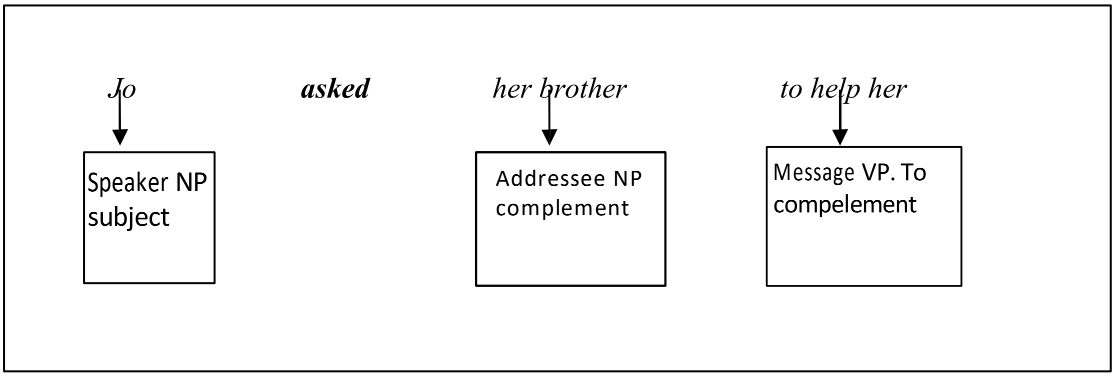


Fig. 4.1 Sentence Annotation for *ask* in Semantic Framework (REQUEST)

The keyword in the sentence above is *ask* which is a member of the order (REQUEST). Each FE and the three pieces of information that describe it (*Jo*, Speaker, NP, subject) are included in one valence group. Thus, there are three groups of valences in the sentence. The three valency groups of the sentence form a valency pattern as follows.

Speaker/NP/subject   Addressee/NP/complement   Message/VP/complement

Fig. 4.2 Valency pattern of *ask* in FE (REQUEST)

The overall pattern of different valencies found in the corpus of an LU is called the valency description. It is very crucial information to include in the entry. The following is the presentation of LU *ask* in Middle English Dictionary (MED).

ask /…/ verb ***

1 [I/T] to speak or write to someone in order to get information from them: *I wondered who had given her the ring but was afraid to ask.* […] **ask (sb) why/how/whether** etc: *She asked me how I knew about it.* **ask (someone) about something**: *Did you ask about the money?* […]

2 [I/T] to speak or write to someone because you want them to give you something: *If you need any help, just ask.* **ask (sb) for sth**: *The children were asking for drinks.* **ask sb's permission/advice/opinion** etc: *I think we'd better ask your mum's opinion first* […]

3 [I/T] to expect someone to do something or give you something: **ask sth (for sth)**: *It's*

a nice house, but they're asking over half a million pounds. […] **ask sb (not) to do sth**:

We ask guests not to smoke in the hotel. […]

4 [I/T] to say that you want something to happen, or that you want someone else to do

something: **ask sb (not) to do sth**: *Then the computer will ask you restart it. He asked us to join him.* **ask to do something**: *I asked to see the manager.* **ask (not) to be**: *The writer has asked not to be named.* **ask that sb (should) do sth**: *The committee has asked that this scheme be stopped for now.*

5 [T] to invite someone to do something or go somewhere with you: **ask sb to sth**: *How many people have you asked to the party?* **ask sb for sth**: *We should ask them for a meal sometime.* […] **ask sb to do sth**: *They asked me to stay the night.*

Fig.4.3. Entry *ask* with valency description

Indonesian verbs are categorized based on their syntactic behavior and divided into six types as follows.

1) transitive verbs with object: *membersihkan* 'to clean', *mencintai* 'to love',*memperlancar* 'to expedite'

2) transitive verbs with an object and complimentary: *menemukan* 'to find', *menuduh* 'to accuse', *mengirimi* 'to send'

3) Semitransitive verbs: *membaca* 'to read', *menulis* 'to write', *minum* 'to drink'

4) intransitive verbs without complementary: *mandi* 'to bath', *bekerja* 'to work', *bertani* 'to farm'

5) intransitive verbs with complementary: *mulai* 'to start', *kedapatan* 'to be found'

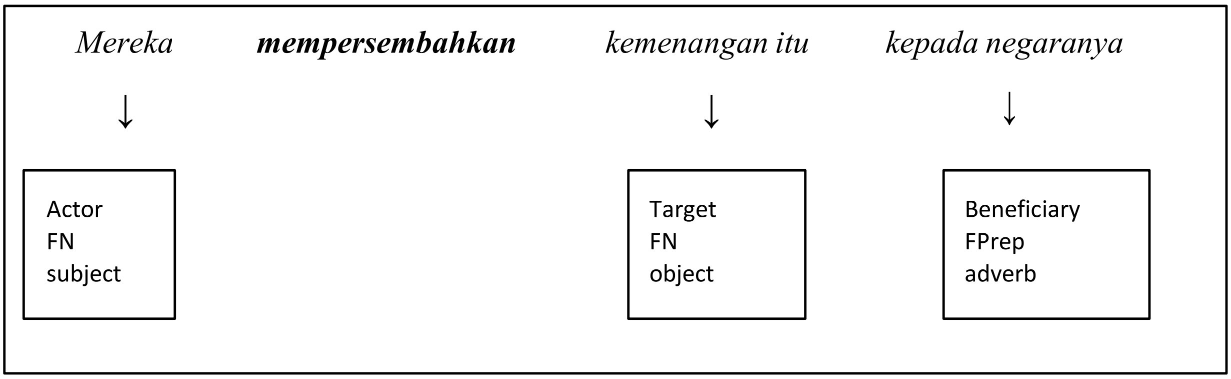
6) Intransitive verbs with noun complimentary and fixed prepositions: *berangkat dari/ke* 'to depart from/to', *menyesal atas* 'to regret for', *sesuai dengan* 'according to'

The division of verbs can also be done with various basic considerations. Kridalaksana (2007) made a subcategory of verbs based on seven criteria, namely (1) the number of accompanying nouns, (2) the relationship of the verb to the noun, (3) the interaction between the accompanying nouns, (4) the argument reference, (5) the identification relationship between the arguments. argument, (6) telis-atelis verb, and (7) performative-constatative verb.

From the learner’s corpus, the verb *mempersembahkan* is obtained. The verb is formed from a gradual process. The premier base *sembah* ‘worship’ received a confix {per-/-kan} that it becomes *persembahkan* ‘to present’. This secondary root then receives the prefix {meng-} to become *mempersembahkan* ‘to present’. Take, for example, the verb *mempersembahkan* in the following sentence.

Mereka mempersembahkan kemenangan itu kepada negaranya ‘
They dedicate the victory to their country’

The verb *mempersembahkan* can be put in frame semantic (**GIVE**). The elements in the sentence can be divided according to their semantic roles into three FE, namely (**Actor**) *mereka* ‘they’, (**Target**) *kemenangan* ‘the victory’, and (**Beneficiary**) *kepada negaranya* ‘to their country’.



Mereka ***mempersembahkan*** *kemenangan itu* *kepada negaranya*

Actor
FN
subject

Target
FN
object

Beneficiary
FPrep
adverb

Fig. 4.4 Sentence annotation of *mempersembahkan*

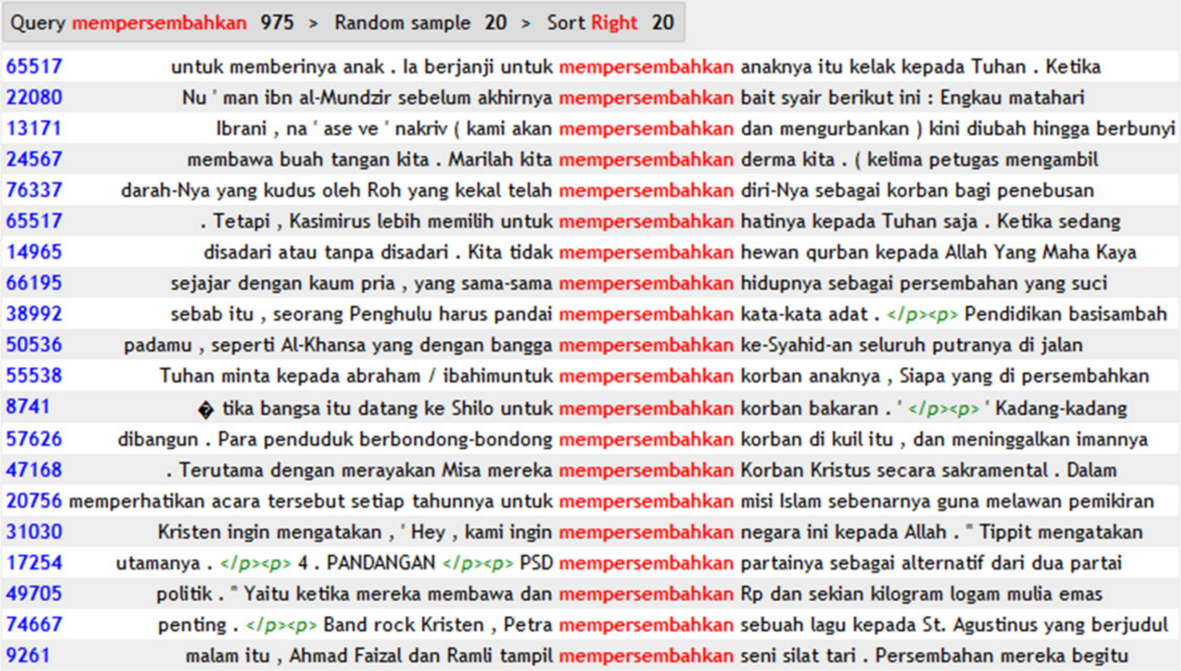
The three FEs above are annotated according to their grammatical functions and syntactic categories. The annotation results for each FE form one valency group (as indicated by the small boxes under each FE) so that in Figure 4.4. there are three valency groups. The first valency group is (*Mereka*, Actors, FN, subject), the second one is (*kemenangan itu*, Target, FN, object), and the third one is (*kepada negaranya*, Beneficiary, FPrep, adverb).

Having each valency group, the FEs are then combined to make a pattern. The resulting valency pattern shows the sentence elements that usually appear together with the word mempersembahkan. The valency pattern is as seen in Figure 4.5 below.

Actor/FN/subject Target/FN/object Beneficiary/FPrep/adverb

Fig. 4.5 Valency pattern of *mempersembahkan*

The valency pattern above serves as a guide for finding the correct collocation in the corpus. From the search results for words using Sketchengine, the calculation of the occurrence rate of the word is 975. From that, there are twenty samples taken. The concordance lines are as follows.



Query mempersembahkan 975 > Random sample 20 > Sort Right 20

65517	untuk memberinya anak . Ia berjanji untuk	mempersembahkan	anaknya itu kelak kepada Tuhan . Ketika
22080	Nu ' man ibn al-Mundzir sebelum akhirnya	mempersembahkan	bait syair berikut ini : Engkau matahari
13171	Ibrani , na ' ase ve ' nakriv (kami akan	mempersembahkan	dan mengurbankan) kini diubah hingga berbunyi
24567	membawa buah tangan kita . Marilah kita	mempersembahkan	derma kita . (kelima petugas mengambil
76337	darah-Nya yang kudus oleh Roh yang kekal telah	mempersembahkan	diri-Nya sebagai korban bagi penebusan
65517	. Tetapi , Kasimirus lebih memilih untuk	mempersembahkan	hatinya kepada Tuhan saja . Ketika sedang
14965	disadari atau tanpa disadari . Kita tidak	mempersembahkan	hewan qurban kepada Allah Yang Maha Kaya
66195	sejajar dengan kaum pria , yang sama-sama	mempersembahkan	hidupnya sebagai persembahan yang suci
38992	sebab itu , seorang Penghulu harus pandai	mempersembahkan	kata-kata adat . </p><p> Pendidikan basisambah
50536	padamu , seperti Al-Khansa yang dengan bangga	mempersembahkan	ke-Syahid-an seluruh putranya di jalan
55538	Tuhan minta kepada abraham / ibahimuntuk	mempersembahkan	korban anaknya , Siapa yang di persembahkan
8741	� tika bangsa itu datang ke Shilo untuk	mempersembahkan	korban bakaran . ' </p><p> ' Kadang-kadang
57626	dibangun . Para penduduk berbondong-bondong	mempersembahkan	korban di kuil itu , dan meninggalkan imannya
47168	. Terutama dengan merayakan Misa mereka	mempersembahkan	Korban Kristus secara sakramental . Dalam
20756	memperhatikan acara tersebut setiap tahunnya untuk	mempersembahkan	misi Islam sebenarnya guna melawan pemikiran
31030	Kristen ingin mengatakan , ' Hey , kami ingin	mempersembahkan	negara ini kepada Allah . " Tippit mengatakan
17254	utamanya . </p><p> 4 . PANDANGAN </p><p> PSD	mempersembahkan	partainya sebagai alternatif dari dua partai
49705	politik . " Yaitu ketika mereka membawa dan	mempersembahkan	Rp dan sekian kilogram logam mulia emas
74667	penting . </p><p> Band rock Kristen , Petra	mempersembahkan	sebuah lagu kepada St. Agustinus yang berjudul
9261	malam itu , Ahmad Faizal dan Ramli tampil	mempersembahkan	seni silat tari . Persembahan mereka begitu

Fig. 4.6 Sketchengine search results view for *mempersembahkan*

From the result display, it appears that there are three possible meanings carried out by **mempersembahkan** based on the words that appear on the right. The first possibility is the meaning contained by the collocation of **mempersembahkan** with words such as *anak* 'child', *hewan* 'animal', *diri* 'self', *hidup* 'life', and *korban* 'sacrifice'. The second possibility is the meaning that arises from the collocation of the word offering with (*syair* 'verse') *puisi* 'poetry', *tarian* 'dance', (*sebuah* 'a') *lagu* 'song', and (*seni* 'art') *tari* 'dance' *silat.* The third possibility is the meaning of the collocation of the word **mempersembahkan** with words such as *derma* 'charity', *misi* 'mission', *negara* 'state', *partai* 'party', and *Rp*. The three groups of words to the right of the word *mempersembahkan* form different meanings. If the verb *mempersembahkan* is followed by words such as *anak*, *hewan*, *diri*, *hidup,* and *korban*, the meaning that appears is 'to sacrifice', while if it is followed by (*bait*) *puisi*, *tarian*, (*sebuah*) *lagu*, and (*seni*) *tari silat* the verb **mempersembahkan** has the meaning of 'to show'. If the **mempersembahkan** is followed by words such as *derma*, *misi*, *negara*, *partai*, and *Rp*, the verb has the meaning of 'to give'.

Based on the grouping of meanings above, the definition for the verb **mempersembahkan** can be formulated according to the order of meaning. The order of the polysemes is arranged according to the meaning that appears first in the sample. The accompanying words for the verb **mempersembahkan** are included in the valency pattern for each polysemy. The valency pattern is not only filled with words in categories, but it is also displayed in the form of lexical words. The formulation of the definition of **mempersembahkan** is as follows.

> **sembah** /səmbah/ *v* [**disembah**, **sembahan**] memberi hormat (biasanya digunakan dalam bentuk pasif): *hanya Tuhan yang patut kita sembah/disembah;* **mempersembahkan** /məmpərsəmbahkan/ *v* [**dipersembahkan**/ **persem- bahkan**, **persembahan]** → **sembah**
>
> 1 **MENGORBANKAN** jika seseorang mempersembahkan kurban berupa anak atau hewan dalam suatu upacara adat, berarti orang tersebut mengorbankan sesuatu untuk sesuatu yang mereka sembah supaya melindungi mereka dari sesuatu yang buruk (**mempersembahkan** + **anak/hewan/diri/hidup/korban** + **kepada+N)**: *mereka mempersembahkan hewan ternak kepada dewa dalam upacara itu*;
>
> 2 **MEMPERTUNJUKKAN** jika seseorang mempersembahkan lagu, tari, syair dll dalam sebuah pertunjukkan, berarti dia mempertunjukkan lagu tersebut untuk

menghibur penonton **(mempersembahkan + syair/tarian/lagu/tari)**: *gadis-gadis itu* *mempersembahkan sebuah tarian di ha-dapan para tamu*;

3 MEMBERIKAN jika seseorang mempersembahkan medali, piala dll berarti dia membe-rikan piala itu sebagai bentuk hormat **(mempersem- bahkan + piala/ medali/derma/ misi/negara/ uang + kepada/bagi/ untuk + N)**: *pemain bulu-tangkis itu mempersembahkan piala kepada negaranya*

This verb has been through semantic framework analysis. The result is a valency pattern that contains the elements that exist together with the verb **mempersembahkan**. In addition, the example entry also provides the collocation for the verb. The order of collocations is arranged according to occurrence in the corpus. Each LU is given a new line and always preceded by a short or reduced definition. The function of bold and capital letters in reduced definitions is also a guide or navigation tool that guides users to find the meaning of LU needed.

Based on the same analysis as **mempersembahkan**, here are other five verb entries that represent each complex affix.

1. {meng-/-kan}: **membersihkan** 'to clean'

bersih /bərsih/ *a* **[bersih-bersih, dibersih-kan, membersihkan] membersihkan** / məmbərsihkan/ *v* [**bersiin** (*inf*)] → **bersih**

1 MENCUCI/MENGGOSOK jika seseorang membersihkan wajahnya, berarti dia mencuci wajahnya dengan air **(~ + wajah/ diri/ tangan)**: *jangan lupa membersihkan ta-ngan sebelum makan*

2 MERAPIKAN jika orang-orang bergotong royong membersihkan lingkungan, berarti mereka merapikan lingkungan tersebut **(~ + meja/kamar/rumah/lingkungan)**

3 MEMBUANG (1) jika seseorang membersihkan saluran, berarti dia membuang semua kotoran yang menyumbat saluran tersebut: *kegiatan akhir pekan ini diisi dengan membersihkan pipa dan saluran yang tersumbat*; **(2)** jika seseorang membersihkan ikan, berarti orang itu membuang semua bagian isi perut ikan itu dan hanya menyisakan dagingnya

2. {meng-/-i}: **mencintai** 'to love'

cinta /cinta/ *n* [**bercinta, bercinta-cintaan, dicinta, dicintai, mencinta, mencintai, tercinta] mencintai** /məncintai/ *v* → **cinta**

1 MENYAYANGI (1) jika seseorang mencintai saudaranya, berarti dia menya-yangi saudaranya tersebut **(~ + saudara/ sesama/orang tua)**: *agama mengajarkan kepada kita untuk mencintai sesama manusia*; **(2)** jika seseorang mencintai pa-sangannya, berarti orang itu mempunyai perasaan yang sangat kuat terhadap pasangannya tersebut **(~ + pasangan/ suami/ istri)**: *dia sangat mencintai pasang- annya dan bersedia menerima segala kekurang-annya*;

2 MENGINGINKAN jika seseorang mencintai perdamaian, berarti dia sangat menginginkan perdamaian tersebut terwujud **(~ + perdamaian/ kedamaian)**: *Indonesia ada-lah negara yang sangat mencintai perda-maian*;

3 MENYUKAI jika seseorang mencintai kebersihan, berarti dai sangat menyukainya **(~ + keindahan/kebersihan)**: *Tuhan itu Mahaindah dan mencintai keindahan*;

4 MENIKMATI jika seseorang mencintai pekerjaannya, berarti dia sangat menikmati saat sedang bekerja **(~ + pekerjaan/hobi)**: *dia tidak pernah bosan karena sangat mencintai pekerjaannya*;

5 MENGHORMATI/BERBAKTI jika seseorang mencintai Tuhannya, berarti dia sangat berbakti kepada Tuhannya dan taat beribadah **(~ + orang tua/Tuhan/ pencip-ta)**: *orang yang mencintai uhannya pasti akan selalu menuruti perintah-Nya*

3. {memper-}: **memperlancar** 'to expedite'

lancar /lancar/ *a* [**dilancarkan**, **melancarkan**, **memperlancar**] **memperlancar** / məmpərlancar/ *v* [**ngelan-carin** (*inf*)] → **lancar**

1 MEMPERMUDAH jika seseorang mem-perlancar suatu urusan, berarti orang tersebut membuat suatu urusan menjadi mudah dan cepat selesai **(~ + urusan/izin)**: *tujuannya hanya untuk membantu memperlancar urusan;*

2 MEMBANTU jika seseorang ingin mem-perlancar tugas orang lain, berarti dia membantu orang lain itu mengerjakan tugasnya supaya selesai **(~ + tugas/pekerjaan)**: *Ibu membantu memperlancar tugas kantor ayah dengan cara mengetikkan surat*

3 MEMPERMUDAH jika dokter memberi-kan obat untuk memperlancar sirkulasi darah, berarti dokter membuat aliran darah menjadi cepat mengalir **(~ + sirkulasi/ alir-n)**: *jus mangga dipercaya dapat mengu-rangi dehidrasi dan memperlancar sirkula-si darah*

4. {memper-/-i}: **memperbaiki** 'to fix'

baik /baik/ *a* [**berbaik-baik, berbaikan, diperbaiki, membaik, membaik, membaiki, membaikkan, memperbaiki] memperbaiki** /məmpərbaiki/ *v* → **baik**

1 MENINGKATKAN jika pemerintah ingin meningkatkan kualitas tenaga kerja, berarti pemerintah ingin kualitas tenaga kerja tersebut menjadi me-ningkat **(~ + kualitas/ mutu/taraf)**: *profesionalisme guru akan memperbaiki kualitas pendidikan*;

2 MEMPERBAGUS jika suatu usaha dilakukan untuk memperbaiki citra seseorang atau sesuatu, usaha tersebut berarti membuat citra tersebut menjadi lebih bagus **(~ + citra/ diri)**: *segala usaha dila-kukan untuk memperbaiki citra dirinya yang sudah terlanjur jelek*

3 MENGUBAH jika seseorang ingin memperbaiki nasib, berarti dia ingin mengubah nasibnya menjadi lebih baik **(~ + nasib)**: *kerja keras menjadi salah satu cara untuk memperbaiki nasib*;

4 MEMBETULKAN jika seseorang ingin memperbaiki suatu kesalahan, berarti dia ingin membe-tulkan kesalahan itu pada kesempatan yang lain **(~ +kerusakan/kesalahan/ kelemahan)**: *pergunakanlah kesempatan yang kedua untuk memperbaiki kesalahan yang dilakukan sebelumnya*

5. {member-/-kan} **memberlakukan** 'to apply'

laku /laku/ n [**berlaku**, **melakukan**, **memberlaku-kan**, **memperlakukan**] **memberlakukan** /məmbərlakukan/ *v* → **laku**

1 MENERAPKAN jika pemerintah memberlaku-kan suatu peraturan, berarti pemerintah menerapkan peraturan tersebut supaya dipatuhi **(~ + hukum/ peraturan/undang-undang)**: *pemerintah mulai memberlakukan peraturan pengendalian dampak lingkungan*;

2 **MENGGUNAKAN** jika sebuah sistem diberlakukan, berarti sistem tersebut digunakan **(~ + sistem/metode/cara)**: *perusahaan itu memberlakukan sistem sidik jari untuk mengatur kehadiran karyawan*

In addition to the valency pattern, in MLD it is also recommended to use other navigation tools to make the lexicographic information easily found. The use of reduced definition with capital letters and bold print arranged downwards is one of the efforts to make the user easily find. Another thing that is also relevant to be discussed in the preparation of MLD entries is the use of illustrations and usage notes.

5 References

Allan, Keith. 2001. *Natural Language Semantics*. Oxford: Blackwell Publishers Ltd.

Amalia, D. 2014. “Formulasi Pendefinisian dan Model Pengentrian Verba dalam Kamus Pemelajar Bahasa Indonesia”. Thesis. Jakarta: Faculty of Culural Studies, University of Indonesia.

Atkins, B. S., dan Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Dzieminanko, A. 2006. *User-Friendliness of Verb Syntax in Pedagogical Dictionary of English*. Tübingen: Max Niemeyer Verlag, Lexicographica Series Maior 130.

Fillmore, C. J. *Charles J. Fillmore Official Website*. Accessed from <http:// linguistics.berkeley.edu /people/fac/fillmore.html> April, 12 2021.

Fontenelle, T. 2008. *Practical Lexicography: The Reader.* Oxford: Oxford University Press. FrameNet. *About FrameNet.* Accesed <http://framenet.icsi.berkeley.edu/> on March 12, 2021.

Kridalaksana, Harimurti. 2007. *Kelas Kata dalam Bahasa Indonesia*. Edisi kedua. Jakarta: PT Gramedia Pustaka Utama.

Madia, I.M., 2001. *Kejutan Pembelajar Asing Menggunakan Kata Berafiks dalam Bahasa Indonesia: Kasus Kata Berafiks ber- dan meng-(kan)*. accessed <http://www.ialf.edu /kipbipa/papers/ Imademadia.htm> on Mei 1, 2021.

Lew, R.. 2004. *Which Dictionary for Whom?: Receptive Use of Bilingual, Monolingual, and Semi-bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex.

Lew, R., dan Dziemianko, A. 2006b. “A New Type of Folk-Inspired Definition in English Monolingual Learner’s Dictionaries and Its Usefulness for Conveying Syntactic Information” in *International Journal of Lexicography* Vol.19 No.3, (hlm.225-242).

Rundell, M. 2006. “Learners’ Dictionaries” in K. Brown (ed.) *Encyclopedia of Language and Linguistics Amsterdam*: Elsevier Ltd. (hlm. 739-743).

Sinclair, J. 2004. *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamin Publishing Company.

Widawati, R. 2002. *Kesalahan Afiksasi dalam Pembelajaran BIPA*. Accessed <http://file.upi.edu/Direktori> on Mei 1, 2021.

POLICY REFORMS IN INDONESIAN VOCABULARY ENRICHMENT PROGRAMME

E. Aminudin Aziz
National Agency for Language Development and Cultivation, Indonesia

Introduction

National agency for language development and cultivation, widely known as Badan Bahasa, is one of nine main units in the Ministry of Education, Culture, Research, and Technology, of the Republic of Indonesia. It has two centres: Centre for Language Cultivation and Centre for Language Development and Preservation. It also has one Secretariat who help with the administration issues. Badan Bahasa has branch offices spread in 30 provinces in Indonesia, supported by 1261 staff all over the provinces.

The main task of Badan Bahasa is first, to develop, cultivate, and preserve the Indonesian language and literature. As part of the development and cultivation, Badan Bahasa also help promote the enhancement of the functions of the Indonesian language to become an international language. The other task of Badan Bahasa, in coordination with the local governments, to help preserve regional languages and literature.

Indonesian and Regional Languages

As a State Language, Indonesian serves as the official language and the national language. As the official language, Indonesian is used

a. In regulation, legislation and official documents,
b. In official speech of President, Vice President, and other state officials who delivered at home or abroad,
c. As language of instruction for education,
d. As language of national-level communication,
e. As means of developing national culture,
f. In transaction and commercial documentation,
g. As means of developing and utilizing knowledge, technology, and art,
h. In mass media,
i. In public administration service,
j. In national or international forum in Indonesia, and
k. For geographic names in Indonesia.

As the national language, Indonesian is

a. national identity and pride
b. a means of unifying various ethnic groups
c. interregional and intercultural communication

Indonesia has 718 regional languages (excluding dialects and subdialects) that identified from 2,560 observation areas of 33 provinces in Indonesia. Those languages are mostly distributed in the eastern part of Indonesia, namely: Nusa Tenggara, Maluku, Maluku Utara, Papua, and some part of Sulawesi. The variation of regional language in the western part is not quite much. More than 2/3 of the regional languages of Indonesia are found in the east part of the country.

Main Issues

Badan Bahasa currently has several language development programme. Those programme are part of Badan Bahasa's mission. There are:

- linguistic and literature research,
- developing teaching materials for the purpose of literacy programme,
- developing language proficiency test or UKBI,
- translation, and
- publication of the works.

Two things that should be highlighted from the programme are vocabulary enrichment and standardization and codification of language system. Vocabulary enrichment is done for adding the great dictionary of Indonesian or KBBI entries.

Some issues regarding entry proposals from regional languages to KBBI are as follows.

- Of 26,345 entry proposals (data per October 2020), there were only 6,044 entry accepted (4.3 %).
- The reason for low acceptance rate is because the entry proposals were not properly selected.
- So far, Badan Bahasa has identified 104 labels of regional languages in the KBBI.
- In average, there were only 58 entries accepted for each languages.

There are specific requirements for new entries to be accommodated in the KBBI:

1. The concept has to be unique and not yet exist in Indonesian.
2. The word must have a high frequency of use by the speakers, either by the speakers of the language or the speakers of other languages that has been exposed to the word. This can be indicated in the corpus.
3. The word has to be euphonic. Easy to pronounce, easy to remember, and also nice to be listened.
4. The word proposed must be able to be formed according to Indonesian morphology rules.
5. The word does not have negative connotation.

The Great Indonesian Monolingual Dictionary or *Kamus Besar Bahasa Indonesia* (KBBI) is intended to be a general dictionary. It is organized to historically record Indonesian usage from time to time, including archaic, obsolete and classic words. The current version, the online format of the dictionary is updated twice a year, which is in April and October. Last month we have just updated the KBBI into the latest version. Other than the online format, there is also printed version of KBBI. However the printed version is not as updated as the online one. The offline format of KBBI also available and can be downloaded via play store for Android and app store for iOS. KBBI is also available for the blind or low vision person in audio format and also in braille. The following table shows the addition of KBBI entries from the first edition to the latest.

Table 1 Number of KBBI's entries

Edition	Year of publication	Number of entries
I	1988	62.000
II	1991	72.000
III	2000	78.000
IV	2008	90.049
V	2018	114,665 (latest update, April 2021)

Previous Policy in regards of vocabulary enrichment

In the past we requested all the branch of the agency in all provinces of Indonesia to contribute massively. They were required to propose entries from their regional languages in the equal amount of target that is 1000 new entries per year per brand office. This was not quite effective and resulted in low acceptance of words proposed to enter KBBI.

One of the main reasons for the ineffective policy is that the linguistic conditions in each region where the branch office is located are not the same. There are areas that have a large number of speakers, but with a small number of languages (for example on the islands of Java and Sumatra). On the other hand, there are areas where there are a lot of regional languages, but very few speakers (for example, the islands of Papua and Maluku). These different linguistic conditions make the same target for each region irrelevant and therefore must be adjusted based on the linguistic conditions. In addition, prior to this, proposals entered into the KBBI were not pre-selected. This resulted in many proposals that did not meet the criteria and was eventually rejected.

The other policy was to use UN official languages as the primary sources for new entries from foreign sources. The problem is, not all UN languages are widely used among Indonesian speakers. Other non-UN official languages are more commonly entered into Indonesian along with the inclusion of popular culture and culinary terms. Languages such as Korean, Japanese, Turkish, and Thai are the languages that influence the Indonesian vocabulary today.

The last was to include all entries from technical terms of specialised dictionaries in KBBI. In fact, not all technical terms can be included in the KBBI because it is a general dictionary. Therefore, every technical term that enters the KBBI must be defined and selected based on its use in general communication.

Current Policy

Improvements must be made to make the process of vocabulary enrichment more effective. For that purpose, we have created a new policy as follows.

1. Knowledge management programme through expert group formation

Currently we have formed ten groups of linguistic expertise and professional service *(Kelompok Kepakaran dan Layanan Profesional, KKLP*). In the group there are people of the same interest and expertise from different work units that are focus more on the programme. Among the groups is KKLP of lexicography and terminology that one of their programme is to focus on the enrichment of the vocabulary.

2. Stratified selection of all proposals for vocabulary enrichment

This is done through several stages, namely: vocabulary inventory, verification in a workshop, and validation in regional language commission meeting. The target is adjusted to the linguistic conditions in each area. This effort is to ensure that proposals submitted to the KBBI are pre-selected so that they are more likely to be accepted. One of the programs in this KKLP is improving the competence of lexicographers. This effort is carried out by holding practical training on lexicography and updating information on lexicography in the form of workshops and seminars.

3. Expanding source languages for enrichment

We focus on influential foreign languages that widely used and have intensive contact with Indonesian: English, Korean, Japan, Turkish, Thai, etc.

4. Digitalisation

Building a database from available language resources (142 printed, 76 electronic, and 6 online dictionaries of regional languages). This effort was carried out mainly to save data in the form of regional language printed dictionaries that did not yet have a digital form. The currently data cannot be fully utilized and processed. We need a big and adequate linguistic database that can be used for lexicography, terminology, and other relevant linguistic purposes.

5. Crowdsourcing system and collaborative work

This is done to ensure all relevant stakeholders can contribute to vocabulary enrichment (universities, mass media, publishers, writers, etc.). For this, we have provide a proposal feature in the KBBI Online application.

Expanding source languages for vocabulary enrichment

There are three language sources for Indonesian vocabulary enrichment: Indonesian language itself, regional language, and foreign language.

Sources from Indonesian language can be found in:

- dictionaries, encyclopaedia, glossaries, thesaurus, corpus of Indonesia (Koin)
- scientific journals
- popular magazines for colloquial and informal vocabulary
- literature work
- official document such as regulation and speech.

Sources from foreign language can be obtained from any widely used, intensively contact, and influential foreign languages such as English, Korean, Japan, Turkish, Thai, etc.

Domain of adoption:

- technology
- pop culture
- culinary
- fashion
- life style, etc.

Sources from regional languages can be obtained from dictionaries of regional languages and linguistic corpora.

Domain of adoption:

- flora and fauna
- marine vocabularies
- traditional architect
- traditional culinary
- local tradition
- culinary, etc.

Language Development Programme

More and more vocabularies from all regional languages in Indonesia will be well-represented in KBBI. To achieve this we had done several efforts as follow.

1. **Platform migration**

We have moved from traditional lexicography to e-lexicography with collaborative workplace, esp. for field lexicography of regional languages.

2. Product diversification

Our product is now user oriented. We provide:

- mobile lexicography for digital natives
- Windows application for digital immigrants
- audio dictionaries for blind or low vision person
- visual dictionaries for hearing impaired person

3. Digitalisation

We have convert the format from print to digital. The contents are also dynamic and real-time since anyone can propose new words or new meanings. They also can proposed to change word's meaning, or deactivate word or meaning.

4. Resource provision

We are now using Indonesian corpus (Koin) as a big data resources of vocabulary enrichment and develop search engine within the corpus. Since its launch in October 2019, Koin has been developing with a target of 5 million words per year. Koin is intended to be a monitor corpus that includes all Indonesian language data. The development team has laid out the design of Koin with the representation of data types in mind. At the initial stage, data were collected from the scientific genre. The following year, data were added from old literary genres. This year the data was collected from the news genre. There are at least 15 million words that will accumulate in Koin by this year.

To speed up the data collection process, we also use the crowdsourcing method and collaborative work from libraries, publisher, and mass media. These three institutions have large data so that they can increase the amount of data currently available in Koin. It is an open-source that can be freely accessed and utilized for scientific and linguistic practical purposes.

SEMANTIC PROSODIES OF *MERAH* AND *BIRU* IN KBBI

Ema Rahardian
Regional Agency for Language in Central Java Province, Indonesia
ema.rahardian@kemdikbud.go.id

Indonesia word *merah* 'red' and *biru* 'blue' may have literal and metaphorical meanings that construct the definition in different senses. In Kamus Besar Bahasa Indonesia (KBBI), the word *merah* and *biru* are described in positive sense and it described without any feature of semantic prosody. Whereas, word *merah* and *biru* may have positive and negative semantic prosodies depend on the context. The phrase *rapor merah*, for instance, can has (1) literally meaning 'a red report' (positive prosody) and (2) metaphorically meaning 'a bad evaluation on the someone or something's performance' (negative prosody). Derivations form of *merah* like *memerah* as well, for instance *mukanya* ***memerah*** *melihat tingkah laku adiknya yang tidak tahu malu.* Word *memerah* here means 'an anger emotion' which has a negative prosody. Otherwise, word *biru* 'blue' has the same tendency with *merah*. The phrase *film biru*, for example, can has (1) literally meaning 'a film that has a blue cover, blue screen, etc.' (positive prosody) and (2) metaphorically meaning 'a film that contains pornographic scenes' (negative prosody). It derivation, such as *membiru* has positive and negative prosodies too, for example (1) *Namun tahun ini diharapkan rapornya* ***membiru*** *dan menghasilkan keuntungan seiring dengan rencana mengoperasikan delapan pesawat* (2) *Pada jendela jendela kusam dan* ***membiru*** *oleh luka lama dan membekas semu*. The first sentence indicate positive prosody because *membiru* here means 'get a good attainment' while the second sentence has a negative prosody because the word *membiru* means 'a bad condition'. This study aims to describe semantic prosodies of Indonesian word *merah* and *biru*. The study employs a corpus linguistic approach from https://corpora.uni-leipzig.de and Sketch Engine Indonesian Web. Hopefully, this research can be able to give some recommendations to the KBBI so the users get the comprehensive meaning.

Key word: semantic prosodies, corpus linguistic, merah, biru

1. Introduction

Color plays an important role in individual's perception of the word (Elliot et al., 2007). The term of color has various meaning depend on the speaker cognitive. The color often relates to the emotion and feeling depend on the speaker perception and cognitive. In Indonesia, the word *putih* 'white' is usually used to describe beauty, nature, and purity. Moreover, the word of color if collocates with certain word will bring up new meaning with different semantic preference. It is also bringing up certain prosody, not only positive but also negative or neutral prosody.

The word *merah* and *biru* both are base color that may have connotation on it used, not only to refer the corresponding color. For example, the word *merah* if collocates with the word *nilai*, brings a negative prosody. Besides that, this collocation also brings both literal and metaphorical meaning depend on its context. See the table below.

Table 1

No.	Left Context	KWIC	Right Context
1	Makanya jangan buru-buru bersedih saat di rapor anak, keponakan, saudara, atau cucu kita bertengger nilai	**merah**	
	So do not rush to get sad when the report cards of our children, nieces, siblings, or grandchildren have a **scores**	**red**	
2	*FSGI beri* ***nilai***	**merah**	*untuk 1 tahun kinerja menteri Nadiem Makarim*
	FSGI gives **scores**	**red**	for 1 year of performance of minister Nadiem Makarim

All the sentences above shown that the word *merah* collocates with the word *nilai* 'score' become *nilai merah* 'red score'. Although they have the same collocation constructions, they have different meaning. Literal meaning shows in the sentence (1) above because the word *merah* modifies the word *nilai* 'score' which is printed in the report. This phrase collocates with the word *rapor* 'report' so that the word *merah* means the score in the report that has a red color. Meanwhile, the word *merah* in the sentence (2) has metaphorical meaning because even it also collocates with the word *nilai,* but the word *merah* here is not refers to red color. The word *merah* in the sentence (2) refers to the poor performance of something. It is because the phrase *nilai merah* collocates with the abstract word 1 year performance. From the context we know that the 1 year performance of minister is not good so that they get a bad evaluation. Therefore, the word *merah* in the sentence (1) has neutral prosody and in the sentence (2) has negative prosody.

This phenomenon has to be noted in dictionary because it describes how some word is used in community. The meaning must be various in time to time depend on the speaker's mindset. This article aims to describe the semantic prosody of the word *merah* and *biru* along with the use of these word in many context with various meanings. This paper perhaps can support the KBBI lexicograph to complete the sense of color in KBBI.

Semantic prosody in corpus linguistics refers to the word tendency to positive or negative polarity (McEnery & Hardie, 2012). Partington said that semantic prosody is the spreading connotational coloring beyond single word boundaries (1998:68). Regarding to Louw, semantic prosody is not merely about connotation. Louw give the argument that semantic prosody refers to a form of meaning which is established through the proximity of a consistent series of collocates. Louw claim that meaning can rub off another word through habitual collocation (2000:50). Louw idea are the same with Xiao & McEnery. They claim that connotation can be collocational or noncollocational whereas semantic prosody can only be collocational (2006:107). It shows that semantic prosody is interaction between the item and its typical collocates. In this case, the item may take on affective meaning even when it is used with other collocates.

The raw data were obtained from https://corpora.uni-leipzig.de and Sketch Engine Indonesian Web. To explore the semantic prosody of the word *merah* and *biru*, I applied the following procedures. First, I extracted the concordance of the word *merah* and *biru* with its derivation of affix *meng-* and presented them by cross tabulation. Then this datum was identified the prosody tendency and also identified its meaning. The data which I assigned in the paper is translated to the English.

2. Results and Discussions

Based on the https://corpora.uni-leipzig.de corpus, the word *merah* and *biru* has many collocation. I tries to display twenty left and right collocations which are the most as follows

Table 2

Token	Left	Right
Merah	warna (157,875), bawang (151,256), berwarna (149,892), kartu (85,328), lampu (60,250), darah (43,052), benang (39,373), cabai (36,867), zona (32,776), bata (29,438), bendera (27,845), plat (27,618), infra (24,830), gula (22,729), jago (20,735), karpet (17,024), daging (10,754), kacang (10,479), anggur (10,462), rapor (9,273)	putih (46,931), muda (30, 247), jambu (17, 151), maroon (13,986), menyala (13,578), marun (13,241), metalik (7,689), padam (6,982), tua (4,725), keriting (4,543), keunguan (2,301), cerah (2,248), terang (2,171), merekah (1,680), membara (1,269), darah (1,354), pekat (800), melambangkan (655), padamlah (634), akibat (485)
Biru	warna (137,424), berwarna (71,644), cetak (31,357), langit (9,247), mengharu (6,357), benua (5,554), bermata (5,124), kemeja (2,559), paus (2,313), jeans (2,232), kampus (1,930), laut (1,823), cahaya (1,792), darah (1,642), sirip (1,603), putih (1,593), baju (1,590), film (1,373), samudra (1,338), rapor (509)	metalik (29,301), tua (20,624), muda (16,366), dongker (5,436), langit (3,932), gelap (3,160), laut (2,004), cerah (1,975), terang (1,396), melambangkan (652), plat (580), bertuliskan (539), lebam (496), transmisi (451), safir (427), keabu-abuan (420), kehitam-hitaman (333), telor (289), nopol (289)

The collocation of the word *merah* and *biru* can have literal and metaphorical meaning. The underlying word in the Table 2 usually have metaphorical meaning depend on its context. From the table we know that metaphorical meaning is occur in the left collocations

a. Collocations and Semantic Prosody of The Word *Merah*

The word *merah* in the corpus has many collocations and it is bringing up the various meanings. Table 3 shows the variation of meanings of the word *merah* without the involvement of any affix.

Table 3

	Meaning	Sample
1.	*n. serupa warna darah*	*Cari angkot berwarna* ***merah*** *nomor B 0 1 jurusan Grogol – Angke.*
	similar with blood color	Look for the **red** city transportation number B 0 1 majoring Grogol – Angke.
2.	*v. terusik*	*Lihat-lihat dulu apa kritikannya, kalau kritikan itu tidak menyakitkan dan tidak membuat telinga saya* ***merah****, bolehlah.*
	disturbed	Take a look at the criticism, if it doesn't hurt and doesn't make my ears **red**, that's fine
3.	*n. gairah, semangat*	*Namun tiba-tiba wajah Panembahan Senapati yang pucat itu menjadi* ***merah*** *kembali*
	passion, spirit	But suddenly Panembahan Senapati's pale face turned **red** again
4.	*n. nafsu*	*Wajah dan tubuhnya merah menyala oleh nafsu*
	lust	**His/her face and body** is red burn with lust
5.	*adj. marah*	*Muka kalian pasti* ***merah*** *karena menganggap bahwa saya sudah menghujat Islam*
	anger	Your faces must be **red** because you think that i have blasphemed Islam

6.	*adj. malu*	*yang aku tahu kala itu, cinta ala monyet diam-diam membuat pipiku bersemu merah merona hanya karena ada tulisan di buku seperti ini "I Love You by R".*
	embarrassed	all I knew at that time, monkey-style love secretly made my cheeks flush red just because there was writing in a book like this "I Love You by R"
7.	*n. prestasi buruk*	*… dijadikan indikator penilaian kalau banyak **komplain** maka akan **merah**.*
	poor performance	… is used as an assessment indicator if there is a lot of complaint it will red
8.	*n. milik pemerintah*	*Untuk itu Ito tidak bosan-bosan untuk mendorongBUMN **merah** ini menjadi perusahaan terbuka.*
	government property	Ito does not get tired of encouraging this **red** BUMN to become a public company.

From the table above, we know that the word *merah* has various meanings, both metaphorical and lexical meaning. The lexical meaning shown in the first sentence because the word *merah* collocates with the word *angkot berwarna* 'city transportation'. Here, the city transportation refers to the transportation in the town that has a red color. So, the word *merah* in the first sentence has lexical meaning.

Meanwhile, the word *merah* in the (2) until (7) have metaphorical meaning. In the 2nd sentence, the word *merah* collocates with the word *telinga.* Here, the word *merah* is not refers to the color of human's ear, but it refers to the emotional condition of somebody. See the sentences below.

a) *Tidak tahu mengapa mereka tidak mau menjenguk? Apa karena **kemarahan** yang **membikin telinga merah**?*
Don't know why they don't want to visit? Is it because **anger** makes your **ears red**?

b) *Malah dibilang eksklusif, tidak progressif, tidak demokratis tidak modern dan masih banyak lagi **bahasa** yang **bikin telinga merah**.*
In fact, they are said to be exclusive, not progressive, undemocratic, not modern, and there are many **languages** that **make red ears**.

c) *Mereka berargumen bahwa di era terbuka ini pemerintah harus **menanggapi kritik** dengan lapang dada, bukan dengan **merah telinga.***
They argue that in this open era the government should **respond gracefully to criticism**, not with **red ears**.

d) *Kita sudah mengenal akan berbagai **pernyataan dan pertanyaan** Bung Asahan seputar peristiwa 1965 yang sering amat **tajam dan kritis**, barangkali membuat **merah telinga** sementara pihak yang tidak setuju.*
We are familiar with Bung Asahan's **statements** and **questions** regarding the 1965 incident which were often very **sharp** and **critical**, perhaps making the **ears red** of those who disagreed with them.

The phrase *telinga merah* or *merah telinga* means terusik 'disturbed' which collocates with the verb *membikin* 'to make'*, bikin* 'make'*, menanggapi* 'to respond'*, membuat* 'to cause' and with the noun *kemarahan* 'anger', *bahasa* 'language', *kritik* 'critics', and *pertanyaan kritis* 'critical question'. From the collocations, it can assume that the word *merah* in the phrase *telinga merah* or *merah telinga* has a negative prosody.

The word *merah* in the (3), (4), (5), and (6) sentences collocate with the word *muka*/*wajah* 'face', *tubuh* 'body', and pipi 'cheek'. It metaphorically means anger, lust, passion, spirit, and embrassed depends on the context. In the sentence (3), the word *merah* has positive prosody because red is used to assign the passion or spirit. Meanwhile, the word *merah* in the sentence (4) and (5) has a negative prosody because it collocates with the word nafsu 'lust' and menghujat 'blasphemed'. The word lust and blasphemed have negative meaning that makes the prosody is negative. Meanwhile, the word *merah* in the sentence (6) has positive prosody because it collocates with the word *cinta* 'love'. It is usually used to describe happy feeling because of love.

As we see the table (1) above, the word *merah* also has metaphorical meaning 'poor performance' as in the sentence (7) and 'government property' as in the sentence (8). The word *merah* in that sentences have negative prosody because they collocate with the word *komplain* 'complaint' as in the sentence (7) that have negative meaning. Meanwhile, in the sentence (8), the word *merah* collocates with the phrase *mendorong menjadi perusahaan terbuka* 'encouraging to become a public company'. The use of the word *mendorong* 'to push' shows that previously the BUMN don't want to be a public company. It leads a negative prosody of the word *merah.*

On the other hand, the word *merah* when collocates with a particular word, forms a phrase that also bring up a metaphorical meaning. Here is 11 collocates of *merah* which combines with the word *rapor* 'report', *plat* 'plate', *nilai* 'score', *karpet* 'carpet', *siaga* 'alert', *garis* 'line', *lampu* 'light', *kelompok* 'community', *politik* 'politics', *golkar* 'golkar', and *kartu* 'card'.

Table 4

	Left Context	KWIC	Right Context	Prosody
1	*Waktu ini, kinerjapemerintahan SBY dankepemimpinan SBYsendiri dipertanyakansecara luas, danmemberinya*	***rapor merah***	*untuk agendapemberantasankorupsinya.*	negative
	Currently, theperformance of the SBYgovernment and his ownleadership is beingwidely questioned, andhas given him	**a red report-card**	for his agenda oferadicatingcorruption.	
2	*Namun, sayangnya,jangankan peralatancanggih, peralatanmanual pun tampaknyakurang mencukupi dibank*	***plat merah***	*yang dibobol penjahat itu.*	negative
	However, unfortunately,sophisticatedequipment, let alonemanual equipmentseems insufficient at the bank	**red plate**	that the crimi nal broke into.	
3	*Ketua DPP PDIPerjuangan FirmanJaya Daeli memberi*	***nilai merah***	*kepada kepadaseluruh menteri yangmasuk ke dalamKabinet IndonesiaBersatu Jilid IIselama 100 harimasa kerjanya*	negative
	Chairman of the PDI-PDPP, Firman JayaDaeli, gave	**red score**	to all ministers whoentered the UnitedIndonesia CabinetVolume II for 100days of their work.	
4	*Kalauinisukses,ituberarti*	***karpet merah***	*untuk generasi berikutnya .*	positive
	If it is success, it means ageneration.	**red carpet**	for the next	
5	*Interpol keluarkanpenangkapan.*	***siaga merah,***	*yakni perintah*	negative
	Interpol has issued a warrant.	**red alert,**	namely an arrest	
6	*Ujung-ujungnya,dari* kemenangan*perang hak siar,adalah dana investasiyang kembali, bahkan laba yang berlipat.*	***garis merah***	*yang bisa kita tarik*	positive
	In the end,the victory of thebroadcasting rightswar, is investmentfunds that return,even multiple profits.	**the red line**	that we can pull from	

7	*Tetapi, Sjakir agaknyamenafsirkan putusanMK itu diartikansebagai*	***lampu merah***	*bagi komisi yangdipimpinnya. Sayakecewa berat karenaMK merestuipembubaran KPKPN,ujarnya.*	negative
	But, Sjakir seemed tointerpret theConstitutional Court'sdecision as a	**red light**	for the commissionhe led. I am verydisappointed becausethe ConstitutionalCourt approved thedissolution of theKPKPN, he said.	
8	*Senjatayangdipergunakan*	***kelompok Merah***	*dalam kerusuhan Poso ternyata dipasok dari ...*	negative
	The weapons used by	**the redcommunity**	during the Poso riots were apparentlysupplied from ...	
9	*Sebabnya ialah karenamereka belummenemukan penjelasanyang tepat tentang lahirdan berdirinyakekuasaan*	***politikmerah***		negative
	The reason is that they have not found a precise explanation for the birth and establishment of … power	**red political**		
10	*Akbar dengan beranimenantang EdiSudradjat dan membuatjarak dengan ABRIsehingga muncul apayang disebut*	***`GolkarMerah'***	*untuk menyebutkekuatan kubu EdiSudradjat yangberhasil dikalahkanAkbar Tanjung*	negative
	Akbar courageouslychallenged EdiSudradjat and made adistance from ABRI sothat what was called	**the red golkar**	emerged to describethe strength of theEdi Sudradjat campthat Akbar Tanjungdefeated.	
11	*Pertamina dapat terkena*	***kartu merah***	*“monopoli” meski ”rakyat” menghendaki subsidi.*	negative
	Pertamina could be hit by	**the red card**	“monopoly” even though the "people" wanted subsidies.	

There are 11 phrases that form a metaphorical meaning, namely *rapor merah* ‘red report’,

plat merah ‘red plate’, *nilai merah* ‘red score’, *karpet merah* ‘red carpet’, *siaga merah* ‘red alert’, *garis merah* ‘red line’, *lampu merah* ‘red sign’, *kelompok merah* ‘red community’, *politik merah* ‘red politics’, *golkar merah* ‘red golkar’, and *kartu merah* ‘red card’. The *rapor merah* has negative prosody because it collocates with the word *kinerja dipertanyakan* ‘the inquiry performance’. It phrase describes something underperforming. It usually used in government, politics and business contexts. Meanwhile, the word *merah* in this phrase denotes a poor condition or poor performance. The meaning of this phrase is same with the phrase *nilai merah* ‘red score’, namely poor performance. When we see this phrase in the wider context, we can assume that it has a negative prosody. See the wider context through the sentence below:

S*ecara keseluruhan PDI Perjuangan melihat bahwa 100 hari kerja kabinet belum membuat kebijakan dan memberikan manfaat yang signifikan terhadap masyarakat.*

Overall, PDI Perjuangan sees that 100 working days, the cabinet has not made a policy and has provided significant benefits to the community.

Here the phrase *red score* is given to the cabinet that has not made a significant policy and benefit to the Indonesian community. It shows that the word *merah* has negative prosody.

The phrase *plat merah* ‘red plate’ usually used in government and politics contexts. This phrase means government property. Meanwhile, the word *merah* here refers to the government. If we see this phrase, without see the context, it seems that the word *merah* has positive prosody. But, if we see it in the context, it shows that the phrase *plat merah* usually used to addressee a government company that is having problem. Unfortunately, the word *merah* here has negative prosody because it refers to the trouble government company.

Negative prosody also occurs in the phrases *siaga merah* because it collocates with the word *perintah penangkapan* ‘arrest command’. The word *merah* here describes the critical condition. The phrase *lampu merah* also has negative prosody because it collocates with the word *kecewa* ‘disappointed*’.* The word *merah* in the phrase *lampu merah* means warning not to continue some works. The phrase *kelompok merah* refers to the rebel community. It shows that the word *merah* associated with insurrection. In the sentence it collocates with the word *kerusuhan* ‘riots’ therefore it has negative prosody. The phrase *politik merah* also has a negative prosody because it collocates with the word *kekuasaan* ‘domination’. The word *merah* here related with the communist. The word *merah* in the phrase *golkar merah* refers to the opposition group in the Golkar. It collocates with the word *menantang* ‘challenge’ so that it bring a negative prosody. As well as the word *merah* in the phrase *kartu merah*. It also has negative prosody because it collocates with the word *monopoli*. The word *merah* in this phrase means warning not to continue some works.

Unlike previous prosody, the word *merah* in the phrase *karpet merah* and *garis merah* has positive prosody. The word *merah* in the phrase *karpet merah* collocates with the word *sukses* ‘success’ and it means honour. The word *merah* in the phrase *garis merah* collocates with the word *tarik* ‘pull’ and it means underlying cause.

The word *merah* has the negative prosody in the achievement, politics, appeal, limit, and emotion senses. Whereas, it has positive prosody in the honorary and sign senses.

The derivation of the word *merah* with affix *men-* only occurs in the emotion sense. It has positive and negative prosody.

	Table 5		
	Meaning	**Alignment Sample**	**Prosody**
1	*v. menjadi merah*	*Matanya pun* ***memerah*** *karena kelelahan*	negative
	turn red	His eyes were red because of fatigue	
2	*v. bergelora dibakar rindu*	*Tubuhnya pun nyaris* ***memerah*** *karena*	positive
	impersonated	His body was almost flushed from being burned longing	
3	*adj. marah*	*Wajah Khalifah langsung* ***memerah*** *pertanda marah*	negative
	anger	The Khalifah’s face immediately flushed, indicating anger	
5	*v. terusik*	*Tidak mudah* ***memerah*** *telinganya ketika dikritik*	negative
	disturbed	His ears is not easy to flush when he is criticized	

Here, the negative prosody is also dominating.

b. Collocations and Semantic Prosody of The Word *Biru*

The word *biru* in the Sketch Engine also have literal and metaphorical meaning. The literal meaning toward the word *biru* can be seen in the sentences below.

a) *Malam itu dia memakai celana jins biru, kaus biru dan*
tangan kanannya memegang sepucuk payung.
That night he wore blue jean pants, **a blue T-shirt** and in his right hand
held an umbrella

b) *Namun, bila si pelanggar menerima maka ia akan menerima* ***lembar biru*** *yang berarti mereka harus ke bank untuk membayar denda untuk kemudian kembali ke kantor polisi*

However, if the offender accepts (it) then he will receive **a blue sheet**, which means they have to go to the bank to pay the fine and then return to the police station.

c) *salah seorang anak yang bertubuh lebih besar memukul temannya sehingga wajahnya menjadi* ***biru lebam***.

one of the bigger boys hit his friend so that his face turned **black and blue**

The word biru in that sentences has literal meaning because it collocates with the word *baju, lembar,* and *lebam*. It describes the blue color of the noun.

The word *biru* also has various meanings with the negative or positive prosody depends on its contexts and collocations. Table 6 shows the various meaning of the word *biru* without any involvement of the affix.

Table 6

	Meaning	Alignment Sample	Prosody
1	n. prestasi bagus	*Begitu BPS mengangkat data yang menggambarkan capaian yang positif, atau* ***biru rapornya****, komentar mereka langsung ...*	positive
	good performance	As soon as BPS raises data that describes positive achievements, or **blue report** cards, their comments are direct	
2	n. modern	*Seperti halnya para ilmuwan di* ***jaman biru*** *yang merobah dunia dengan teknologi, anak indigo akan merombak dunia dengan terlebih dahulu menata spiritual*	positive
	modern	Just like scientists in the **blue age** who changed the world with technology, indigo children will overhaul the world by first arranging the spiritual	
3	n. ningrat	*Keduanya sama-sama keturunan* ***darah biru*** *Sultan Trenggana dari Demak Bintara.*	positive
	noble	Both of them are descendants of the **blue blood** of Sultan Trenggana from Demak Bintara	
4	n. porno	*pasti bukan* ***film biru****, karena indonesia sudah mempunyai undang-undang anti porno.*	negative
	porn	definitely it is not a **blue film**, because Indonesia already has an anti-porn law.	
5	n. romansa	*Pemilihan bahasanya pun lugas, tidak bertaburan kata puitis* ***mendayu biru*** *.*	negative
	romance	The choice of language is straightforward, not sprinkled with the **blue alluring** poetic words	

6	*v. membanggakan*	*... mengibarkan bendera Merah Putih di kejuaraan antarbangsa . Saat itu , langit bulan September betul- betul terasa* ***biru*** *bagi rakyat Indonesia.*	positive
	glory	… flaping the Red and White flag in the international championship. At that time, the sky in September felt **blue** for the Indonesian people.	
7	*n. spiritual*	*Munculnya Species Manusia Baru Fenomena "****Bocah Biru****" belakangan ini menjadi perhatian para ilmuwan di Rusia.*	positive
	spiritual	The Emergence of a New Human Species The "**Blue Boy**" phenomenon has recently caught the attention of scientists in Russia.	

The word biru in the sentence (1) collocates with the word rapor 'report' that means good performance. This phrase usually uses to address the institution, company, or government that has a good achievement and performance. Therefore, it brings positive prosody. The positive prosody also occurs in the sentence (2). Here, the word biru is collocates with the noun jaman 'era' and the verb merubah 'changes'. The word biru here means modern. The word biru in the sentence (3) likewise. It also has positive prosody because the word biru collocates with the noun darah 'blood', keturunan 'descendants', and Sultan 'king'. The word biru here refers to the noble.

Meanwhile, the word biru in the sentence (4) is not the same. Here, the word biru has negative prosody because it collocates with the word film and porn. The word biru in that sentence means porn. The phrase film biru used to call a film that has pornography content. The word biru in the sentence (5) also has negative prosody because it collocates with the negative word tidak. The word biru in the sentence (4) and (5) have romance sense.

The word biru in the sentences (6) and (7) have positive prosody. It is because the word biru in the sentence (6) collocates with the word kejuaraan. In here, the word biru means glory. The prosody of the word biru in the sentence (7) can be seen in the wider context, as follows.

> *Para ilmuwan mengatakan mereka* ***memiliki*** *kekuatan* ***supernormal****,*
>
> *dapat* ***melihat fenomena ganjil****, dan dapat* ***meramal peristiwa*** *yang akan terjadi. Ciri khas mereka adalah berinteligensi tinggi, berintuisi tinggi, sangat sensitif dan lain-lain.*
>
> Scientists say they have **supernormal powers**, can **see bizarre phenomena**, and **can predict events** that will occur. Their characteristics are high intelligence, high intuition, very sensitive and others

From the context we know that the word *biru* collocates with the verb *memiliki, melihat,* and *meramal* also collocates with the noun *kekuatan supernormal, fenomena ganjil,* and *peristiwa yang akan terjadi.* It shows that the word *biru* has positive prosody because it refers to something supernatural. The word *biru* has positive prosody in the achievement, family, and period.

The derivation of the word *biru* with the affix *meng*- also has various meanings as follows.

Table 7

	Meaning	Alignment Sample	Prosody
1	*adj. melankolis*	*Puncak dari semua itu adalah, malam dimana gw kembali meringkuk dalam kesepian yang* ***membiru*** *dan bisu.*	negative
	melancholy	The culmination of all that was, the night when I huddled again in blue loneliness and mute	
2	*v. menyakitkan*	*Saat yang kubawa dari masa lalu hanyalah luka yang* ***membiru****.*	negative
	painful	The moment I carried from the past was just a wound that turned blue	
3	*v. bernostalgia*	*Nostalgia SMA dirasakan semakin* ***membiru****, bila kita simak lagu di bawah ini.*	negative
	feeling nostalgic	High school nostalgia is felt to be getting more blue, if we look at the songs in SMA Nostalgia, it feels even more blue, if we look at the songs below, SMA nostalgia feels even more blue, if we watch the song below.	
4	*n. kinerja bagus*	*Namun tahun ini diharapkan rapornya* ***membiru*** *dan menghasilkan keuntungan seiring dengan rencana mengoperasikan delapan pesawat.*	positive
	good performance	However, this year it is expected that the report cards will turn blue and turn a profit along with the plan to operate eight aircraft.	
5	*n. kinerja bagus*	*Selain itu, melempemnya kinerja anak perusahaan juga menjadi faktor sulitnya perusahaan membuat kinerja saham* ***membiru****.*	positive
	good performance	In addition, the sluggish performance of the subsidiaries is also a factor in the company's difficulty in making stock performance turn blue.	
6	*v. bernostalgia*	*Ingin rasanya Cunda mengulur waktu ketika merasai hatinya masih* ***membiru.***	negative
	feeling nostalgic	Cunda wanted to buy time when he felt that his heart was still blue.	
7	adj. *melankolis*	*Sendu yang kelabu tak akan berubah warna kecuali semakin* ***membiru****.*	negative
	melancholy	The gray blister will not change color unless it gets more and more blue.	
8	*adj. tak bergairah*	*Suami-istrilah yang bertanggung jawab untuk memelihara, memupuk dan menyiram kasih di antara mereka, agar jangan sampai "beku dan* ***membiru****".*	negative
	low of spirit	Husband and wife are responsible for nurturing, cultivating and watering the love between them, so as not to "freeze and turn blue".	

In the table 7 shown that there are two senses that form a meaning of *membiru*. There are emotion and achievement senses. The negative prosody occurs in the emotion sense as we see in the sentences (1), (2), (3), (6), (7), and (8) meanwhile the positive prosody occurs in the achievement sense like in the sentences (4) and (5).

3. Conclusions

There are various meaning of the word *merah* and *biru* with its derivations. It meaning occurs in different senses that can be added in the KBBI. The senses of the word *merah* refer to the politics, emotions, achievement, appeal, and limit. The politics sense brings various meanings, such as 'communist', 'opposition', 'domination', and 'insurrection'; the emotions sense brings a 'offended', 'anger', 'passion', 'spirit', 'lust', and 'embrassed' meanings; the achievement sense brings a 'bad performance' meaning; the appeal sense brings a 'caution', 'warning', and 'command' meanings; the limit sense bring a 'critical limit' meaning; the honorary sense brings a 'honour' meaning; and the sign sense brings a 'underlying cause' meaning. Its derivational also has the same. The word *memerah* has emotional sense that bring a 'anger' and 'impassioned' meanings.

Meanwhile, the word *biru* has opposite connotation with the word *merah* in the sense of achievement. The sense of the word *biru* are emotion, achievement, family, and period. The emotion sense brings a 'romance' and 'porn' meanings; the achievement sense brings a 'good performance' meaning; the family sense brings a 'noble' meaning; and the period sense bring a 'modern era' meaning. The word *membiru* has emotion and achievement senses. The emotion sense brings a 'sad emotion', 'melancholy', 'hurt feeling', 'nostalgic feeling', and 'low of spirit' meanings. While the achievement sense brings a 'good performance' meaning. Perhaps, that senses can to complete the sense that exist in the KBBI right now. Therefore, the meaning of the word *merah* and *biru* can be more comprehensive.

Bibliography

Elliot, A.J., Maier, M.A., Moller, A.C., Friedman, R., & Meinhardt, J. (2007). "Color and Psychological Functioning: The Effect of Red on Perfomance Attainment". *Journal of Experimental Psychology: General*, 136 (1), 154—168. https://doi.org/10.1037/0096-3445.136.1.154

Goddard, Cliff and Anna Wierzbicka. (2014). *Words and Meaning: Lexical Semantics across Domains, Languages, and Cultures*. United Kingdom: Oxford University.

Louw, Bill. (2000). "Contextual Prosodic Theory: Bringing Semantic Prosodies to Life" in Heffer, Chris, and Helen Sauntson (eds.). *Words in Context, A Tribute to John Sinclair on his Retirement*.

Mc Enery, T. & Hardie, A. (2012). *Corpus Lingustics: Method, Theory, and Practice.* Cambridge: University Press.

Partington, A. (1998). *Patters and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins Publishing.

Philip, Gill. (2011). *Colouring Meaning: Collocation and Connotation in Figurative Language*. Amsterdam: John Benjamins Publishing.

Setiawan, Aan, Totok Suhardijanto, and Setiawati Darmojuwono. (2019). "Semantic Preferences of Indonesia Words Hitam and Putih". *Proceeding of The 4th International Seminar on Linguistics*. pp.1—12. Padang: Andalas University.

Stewart, Dominic. (2010). *Semantic Prosody: A Critical Evaluation*. New York: Routledge.

Stubbs, Michael. (2016). "Corpus Semantics" in Riemer, Nick (ed.). *The Routledge Handbook of Semantics*. New York: Routledge.

Xiao Richard & Mc Enery, T. (2006). "Collocation, Semantis Prosody, and Near Synonymy: A Cross-Linguistic Perspective". *Applied Linguistics* 27 (1): 103—129. https://doi.org/10.1093/applin/ami045

PICTURING DEICTIC TIME WORDS OF BAHASA INDONESIA CORPORA COLLECTION LEIPZIG UNIVERSITY

Eva Tuckyta Sari Sujatna, Larasati Puspa Martani Sugianto
Department of Linguistics, Faculty of Cultural Sciences, Universitas Padjadjaran, Indonesia;
University of Melbourne, Australia

Abstract

Bahasa Indonesia is the official language in Republic of Indonesia. There is still limited research on Bahasa Indonesia related to deictic time words. This paper tries to explore the Bahasa Indonesia deictic time words: *selumbari, kemarin, hari ini, besok* or *esok, lusa, tulat*, and *tubin*. The data concerning the seven Bahasa Indonesia deictic time words were obtained from Bahasa Indonesia corpora collection Leipzig University 2008-2018. There are two research aims: firstly, to categorize the data based on the distribution of each deictic time words and secondly to indicate the pair of each deictic time word and its context in Bahasa Indonesia corpora collection Leipzig University related to the data collected in ten years (2008-2018 except 2014). It was reported, there are only four Bahasa Indonesia deictic time words ("*kemarin*", "*hari ini*", "*besok/ esok*",and "*lusa*") are found in the data while "*selumbar*i", "*tulat*", and "*tubin*" are not. The deictic time words "*besok*" and "*esok*" has similar meaning and they could be paired with "*hari*" to become "*besok/ esok hari*", "*pagi*" to become "*besok/ esok pagi*", *-nya* to become "*besoknya/ esoknya*", "*harinya*" to become "*besok/ esok harinya*", and the combination of the days' name to become "*Senin/ Selasa/ Rabu/ Kamis/ Jumat/ Sabtu/ Minggu besok/ esok*".

Keywords: deictic time words; Bahasa Indonesia; corpora collection

I. Introduction

Many researches on Bahasa Indonesia, but it is still limited research on deictic time words in Bahasa Indonesia. Bahasa Indonesia or sometimes called as Indonesian language or Indonesian is the official language used in Republic of Indonesia since October 28, 1928. Republic of Indonesia or Indonesia is known as an archipelago country. Indonesia has around

17,504 islands and it has about 718 local languages in 34 provinces. To communicate one to another, the Indonesian people speak Bahasa Indonesia as the language of education and the language to unify ethnic languages in Indonesia (Kwary, 2019).

It was reported by National Agency for Language Development and Cultivation (*Badan Pengembangan dan Pembinaan Bahasa*) Ministry of Education and Culture of the Republic of Indonesia, that Bahasa Indonesia has 17,000 vocabularies which are found in *Kamus Besar Bahasa Indonesia* (KBBI). It is a dictionary of Bahasa Indonesia which was published by *Badan Pengembangan dan Pembinaan Bahasa*. It was reported that some of the vocabularies of Bahasa Indonesia were taken from foreign languages, such as Dutch, Japanese, Portuguese, Spanish, French, and English the countries that colonized Indonesia and also Arabic, Chinese, and India as the countries that have close relationship to Indonesia. Not only languages from other countries but also local languages in Indonesia increase the number of Bahasa Indonesia vocabularies, such as Javanese, Sundanese, and Minangkabau language (Nurlina, 2014).

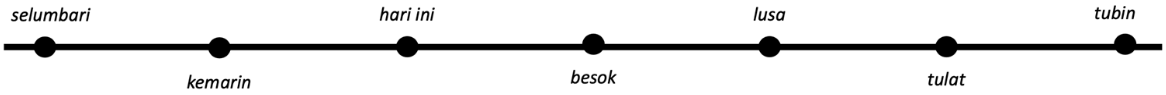
The phrase of deictic time words in this paper is borrowed from Tillman et al. (Tillman, Marghetis, Barner, & Srinivasan, 2016). The term refers to the time, in English such as, yesterday, today, and tomorrow. The fact that every language has different way in expressing time in each language, it includes Bahasa Indonesia. This paper is trying to figure out the use of the deictic time words in Bahasa Indonesia clauses which were collected from Indonesia corpus from corpora collection Leipzig University, from 2008 up to

2018. The corpora collection provided by Leipzig University is not only Bahasa Indonesia, it also provides corpus from many languages around the world. It was reported, in 2007, it coveraged about 20 languages (Biemann, Heyer, Quasthoff, & Richter, 2007) and today it is about 200 languages (Goldhahn, Eckart, & Quasthoff, 2012). It was indicated the corpora collection collected 74,329,815 Bahasa Indonesia clauses which consists of 7,964,109 words types and 1,206,281,985 tokens, so it is identified as the newest and the biggest (Kwary, 2019). Since the corpora collection Leipzig University has a huge number of data, the present writers employ it as the source of the data.

Several researchers did their research on Corpus Linguistics related to Contemporary of Contemporary of American English (Altohami & Salama, 2019; Rafatbakhsh & Ahmadi, 2019; M. L. Sujatna, Sujatna, & Pamungkas, 2019; Yusu, 2014). Some researches on Corpus Linguistics related to Bahasa Indonesia are done, some of the are "A corpus platform of Indonesian academic language" (Kwary, 2019), "Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets" (Wicaksono, Vania, Distiawan, & Adriani, 2014), and "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus" (Larasati, Kubon, & Zeman, 2011).

According to Yuliawati et al. (2019) and supported by Sujatna et al. (2020) that corpus could be connected an empirical research examining the actual patterns of use in natural texts (E. T. S. Sujatna, Heriyanto, Krisnawati, Amalia, & Pamungkas, 2020; Yuliawati, Dienaputra, Sujatna, Suryadimulya, & Lukman, 2019). Earlier Jones and Walter stated that searchable collection of texts electronically stored while the text could be spoken or written in various length, although generally, it would be longer than a single utterance or a single written clause (Jones & Waller, 2015). The earlier argumentation supported the Baker's opinion that corpus linguistics enables the researchers' cognitive and social biases to be reduced due to the fact that none is impossible to claim to be absolutely objective about a piece of research (Baker, 1996).

In Bahasa Indonesia, to indicate deictic time words that refers to the days it could be expressed by *selumbari, kemarin, hari ini, besok or esok, lusa, tulat*, and *tubin* as described in the following picture.



Picture 1

Deictic time words in Bahasa Indonesia

In English, the word "*selumbari*" means 'two days before', "*kemarin*" means 'yesterday', the words "*hari ini*" means 'today', "*besok*" or "esok" means 'tomorrow', "*lusa*" means 'a day after tomorrow' or 'next two days', "*tulat*" means 'two days after today' or 'next three days', and *tubin* means 'three days after today' or 'next four days.' The seven words mentioned are categorized into nouns as stated in KBBI. It was reported that two of the seven deictic time words are borrowed from other languages; the word "*selumbari*" was borrowed from Bahasa Minangkabau as one of the ethnic languages in Indonesia and the word "*tubin*" from *klasik Malayu* or Malay language.

Besides as 'yesterday', the word "*kemarin*" also has different meaning. It was reported, the word "*kemarin*" could be meant '*yang lalu'* or 'last' and it also could be repeated as "*kemarin-kemarin*" means 'several days ago'.

II. Research Method

The research methods applied in this research is the combination of qualitative and quantitative methods. They are combined to produce a comprehend of the topic of the research that either the approach alone (Creswell & Creswell, 2018).

The first procedure done was the present writers collected the data as the corpus from corpora collection Leipzig University. The corpora collection Leipzig University provides the Bahasa Indonesia corpus since 2008-2018 except 2014 which was downloaded from https://wortschatz.uni-leipzig.de/en/download/indonesian#ind_mixed_2013. The corpora were taken from several sources which were categorized into blogs, news, newscrawl, web, web public, and Wikipedia. From the collected data 2008-2018, it was obtained 7,309,185 corpora in Bahasa Indonesia clauses. For the composition of the corpora, the present writers describe it in table 1.

No	**Year**	**Sources**	**Total**
1	2008	news	491,099
2	2009	news	493,535
3	2010	news	249,344
4	2011	news, newscrawl, web	1,504,170
5	2012	blogs, news, newscrawl, web	1,936,880
6	2013	web	512,245
7	2015	newscrawl, web	754,935
8	2016	wikipedia	627,356
9	2017	web, web public	491,680
10	2018	web	247,941
		Total of corpora	**7,309,185**

Table 1

Total of Corpora 2008-2018 (except 2014)

Secondly, the present writers identified the corpora into seven different categories: *selumbari, kemarin, hari ini, besok* (and *esok*)*, lusa, tulat,* and *tubin.* The present writers reported that

29,884 corpora consisting the word *kemarin*, 30,614 corpora found that containing the words "*hari ini*", 11,615 corpora containing the words "*besok*" and "*esok*", 513 corpora consisting the word "*lusa*", and no corpora containing the word "*selumbari*", "*tulat*" or "*tubin*" in corpora collection .

After classifying the corpora, as the next procedure, the present writers analyzed and described the Bahasa Indonesia corpora related to deictic time word. Finally, the present writers outlined the usage of the words "*kemarin*", "*hari ini*", "*besok*" and "*esok*", and "*lusa*" in Bahasa Indonesia corpora found in corpora collection Leipzig University.

III. Result and Discussion

As the aims of the research that the present writers decided, firstly, to categorize the data based on the distribution of each deictic time words related to the data collected in ten years (2008-2018 except 2014). Secondly, the present writers decided to indicate the pair of each deictic time word and its context in Bahasa Indonesia corpora collection Leipzig University related to the data collected in ten years (2008-2018 except 2014). From the collected data the present writers describe four distributions ("*kemarin*", "*hari ini*", "*besok*", and "*lusa*") since there were not found the usage of the words "*selumbari*", "*tulat*", and "*tubin*" in corpora collection Leipzig University, as described in the following chart. It was reported that the present writers found six Bahasa Indonesia clauses containing the word "tubin" but they were excluded from the data collection since they do not refer to the deictic time words in Bahasa Indonesia.

Bahasa Indonesia deictic time word	**No of Data**	**Percentage**
selumbari	0	0
kemarin	29,884	0.408%
hari ini	30,614	0.418%
besok and esok	11,615	0.160%
lusa	513	0.007%
tulat	0	0
tubin	0	0
Total of corpora with deictic time words	72,626	0.010%
Total of corpora without deictic time words	7,236,559	0.990%
Total	**7,309,185**	**100%**

Table 2

Bahasa Indonesia deictic time words in Corpora Collection Leipzig University (2008-2018)

The table 2 explains that 7,309,185 Bahasa Indonesia clauses found in corpora collection Leipzig University (2008 - 2018) and it was only 72,626 or 0.010% Bahasa Indonesia clauses containing deictic time words "*kemarin*", "*hari ini*", "*besok*", or "*lusa*". It means that 7,236,559 or 0.990% Bahasa Indonesia clauses do not consist of deictic time word of each.

3.1 The Distribution of the word "*kemarin*" in Corpora Collection Leipzig University

Based on the corpora collection which was collected, the present writers found 29,884 Bahasa Indonesia clauses containing the word "*kemarin*". Based on KBBI "*kemarin*" /*ke.ma.rin*/ is a noun; it means '(*satu) hari sebelum hari ini*', in English means 'a day before today' or 'yesterday'. From the collected data, it was reported that the word *kemarin* were taken from various sources such as blogs, news, newscrawl, web, web public, and Wikipedia for ten years, since 2008 up to 2018 (excluding 2014) as described in the following diagram 1.

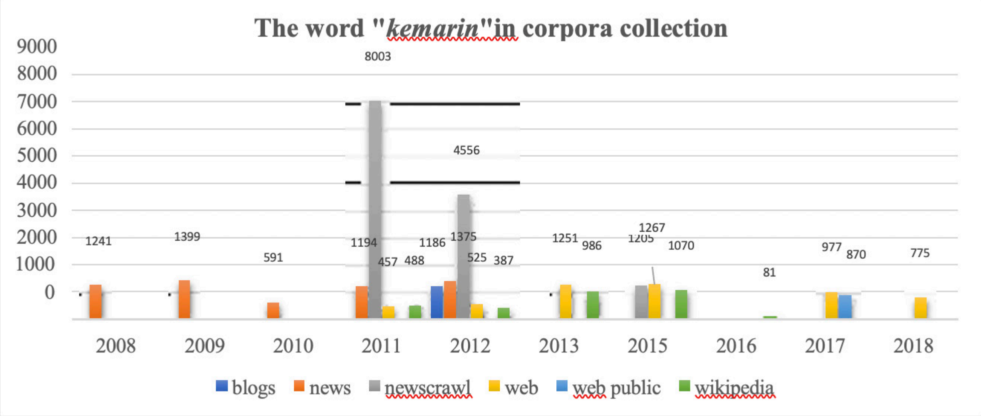


Diagram 1

The word "*kemarin*" in Bahasa Indonesia Corpora Collection Leipzig University (2008-2018)

From the data collected, the word "*kemarin*" was the most frequently used in 2011. It was reported that the number of the corpora "*kemarin*" used in 2011 (the highest number) is 10,142 from the various sources, such as news, newscrawl, web, and Wikipedia. While the lowest frequently used of the word "*kemarin*" found in 2016, it was just only 81 corpora collected from Wikipedia. Besides the meaning of '(*satu*) *hari sebelum* *hari ini*' 'a day before today' or 'yesterday', the present writers found another meaning as described in the following.

a. "*Kemarin*" means '*yang lalu*' or 'last'

It was reported that the present writers found the word "*kemarin*" that has different meaning; it means '*yang lalu*' or 'last' as in "*kemarin September*" 'last September' as illustrated in the following

(1) *Bahkan, karena khawatir pada status induk perusahaan keuangan Fannnie Me dan Freddie, pemerintah memaksa diri untuk mengambil alih eksposur kredit perumahan pada **September kemarin**.*

(2) *ASII pimpin indeks turun setelah JP Morgan melakukan `downgrade`, yang kemungkinan terkait penurunan penjualan **Agustus kemarin**, kata Analis Riset PT Recapital Securities Poltak Hotradero, kepada ANTARA News di Jakarta, Kamis*.

(3) D*ana tambahan diperlukan untuk menambah suntikan 25 milyar dollar yang telah dilakukan departemen keuangan Amerika **Oktober kemarin***.

(4) *Sebelumnya, PT Indonesia Tower melakukan ujicoba penggunaan perangkat teknologi Wimax di hadapan Menkominfo dan Dirjen Postel Basuki Yusuf Iskandar pada **April kemarin***.

The data (1) – (4) "*September kemarin*", "*Agustus kemarin*", "*Oktober kemarin*", and "*April kemarin*" were described that the word "*kemarin*" could be combined with September, Agustus, Oktober, and April as the names of month. The meaning "*kemarin*" in the four but sentences are not 'yesterday' but it refers to 'last' as an adjective. The combination of the words, in English, become 'last September', 'last August', 'last October', and 'last April'.

Besides combining with the names of month, the word "*kemarin*" is usually combined with the names of day as described in the data (5) - (6).

(5) *Ia melihat beberapa pelaku pasar domestik kembali melakukan aksi beli pada beberapa saham unggulan, sehingga mendorong indeks LQ45 mengalami penguatan di akhir sesi perdagangan **Senin kemarin***.

(6) *Peristiwa terbaru yaitu penemuan sebuah bom rakitan di depan rumah I Made Santi, eks transmigran asal Bali yang bermukim di Dusun Mauro, Desa Kawende, Kecamatan Poso Pesisir Utara, pada **Jumat kemarin*** (31/10).

The two data above explained that the word "*kemarin*", as an adjective, could follow the word *Senin* ("*Senin kemarin*" 'last Monday') and *Jumat* ("*Jumat kemarin*" 'last Friday') as the names of day.

Not only the names of month and day could be combined with the word last, but also the word "year" as illustrated in the following data.

(7) *Upaya ini terus kita lakukan sejak **minggu kemarin** sampai beberapa bulan mendatang.*

(8) *Sejak pemeriksaan hari Kamis **pekan kemarin**, bapak menetap di Jakarta di kawasan Jakarta Selatan, di salah satu saudara ibu, ujarnya*.

(9) *Nanti kita lihat secara total berapa pertumbuhan ekonomi **satu tahun kemarin***.

(10) *Kosovo akan menjadi "masyarakat yang menghormati martabat manusia" dan akan menghadapi "warisan menyakitkan dari **masa kemarin**, dalam semangat rekonsiliasi dan pengampunan".*

b. "*Kemarin-kemarin*" means '*beberapa hari yang lalu*' or 'several days ago'

In the corpora collection, the present writers found that the word *kemarin* could be reduplicated to become “*kemarin-kemarin*”; it means ‘*beberapa hari yang lalu*’ or ‘several days ago’ as described in the following.

(11) *Undangan sudah dikirim sejak **kemarin-kemarin** dan sudah diterima karena ada tanda terimanya, kita mengirim utusan khusus untuk membawa undangan itu, ujarnya.*

(12) *Kan, alat penyadapnya baru didengarkan **kemarin-kemarin**," katanya.*

Data (11) and (12) describe the two sentences containing “*kemarin-kemarin*” that function as an adverb. Besides referring to several days ago, the words “*kemarin-kemarin*” which was preceded by the word “*dari*” could have different meaning as in data (13).

(13) *Saat dimintai komentar mengenai kemenangan yang diraih Zheng/Gao setelah lawan mereka mundur, Nova mengatakan, “Itu biasa Mbak, taktik mereka, **dari kemarin-kemarin** begitu melulu*”.

The data (13) describes that “*dari kemarin-kemarin*” refers to ‘*dulu*’ or ‘formerly; previously’ that functioned as an adverbial.

3.2 The Distribution of the words “*hari ini*” in Bahasa Indonesia Corpora Collection Leipzig University

On basis of the corpora collection, the number of the words “*hari ini*” found 30,614 clauses. The words “*hari ini*” means ‘today’ and the corpora obtained were taken from various sources such as blogs, news, newscrawl, web, web public, and Wikipedia for ten years, since 2008 up to 2018 (excluding 2014).

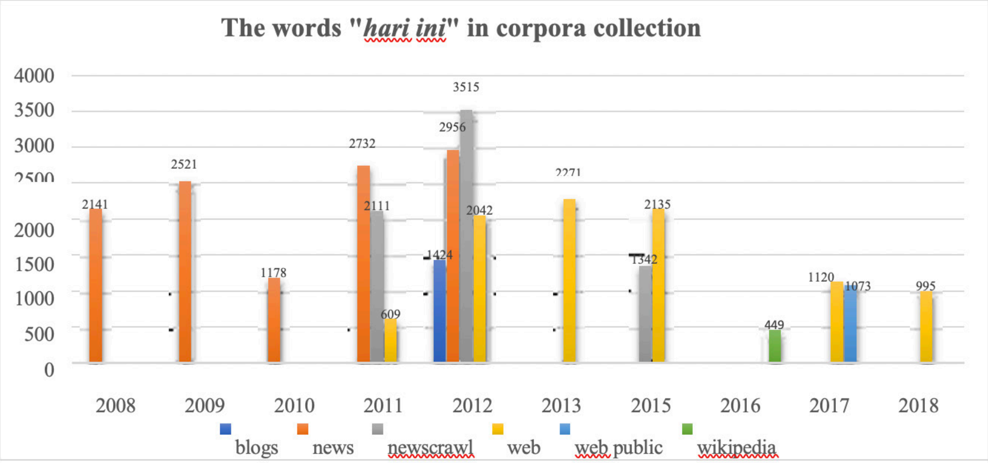


Diagram 2

The words “*hari ini*” in Bahasa Indonesia Corpora Collection Leipzig University (2008-2018)

The data described that the highest number of the words “*hari ini*” usage found in the corpora in 2012; it was 9,935 corpora. The lowest number of the usage of “*hari ini*” was reported in 2016; it was found 449 corpora in Wikipedia.

The use of “*hari ini*” found in corpora collection refer to the similar meaning that it ‘today’ as described in the following data.

(14) ***Hari ini** diputuskan Taufik dan Markis/Hendra tidak berangkat ke Singapura, ujar ketua Bidang Pembinaan Prestasi PB PBSI Lius Pongoh di Jakarta, Jumat.*

(15) *Mereka sudah resmi bertugas **hari ini**, kata Presiden Yudhoyono sebelum memulai Sidang Kabinet mengenai Persiapan Lebaran di Kantor Kepresidenan, Jakarta, Selasa.*

From the collected corpora, it was reported that the "hari ini" refers to 'today' as both data above (14) and (15) described.

3.3 The Distribution of the words "*besok*" and "*esok*" in Corpora Collection Leipzig University

The word "*besok*" in KBBI /*be.sok*/ /*bèsok*/ is a noun. In Bahasa Indonesia, the word "*besok*" has a synonym, it is "*esok*" and it is also a noun. Both of the words "*besok"* and "*esok*" have similar meaning '*hari sesudah hari ini*' while in English it means 'a day after today or tomorrow'. It was reported, the present writers found 9,409 words of "*besok*" and 2206 words of "*esok*" so the total is 11,615 Bahasa Indonesia clauses in corpora collection. The following diagram 3 describes the combination of the words "*besok*" and "*esok*" in Bahasa Indonesia clauses found in the corpora collection Leipzig University in 2008-2018 (except 2014).

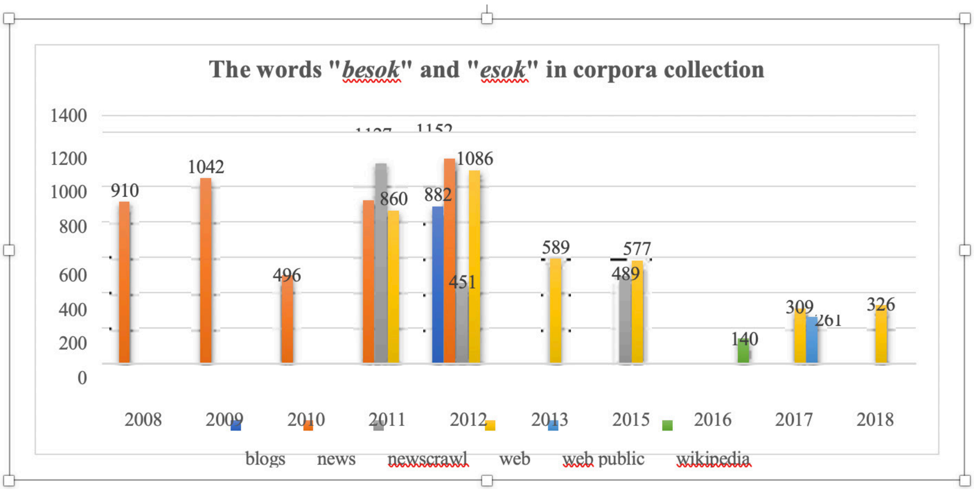


Diagram 3

The words "*besok*" and "*esok*" in Bahasa Indonesia Corpora Collection Leipzig University (2008-2018)

The diagram 3 describes the combination of the words "*besok*" and "*esok*" in Corpora Collection Leipzig University (2008-2018). From the collected data, it was reported that the usage of the word "*besok*" is higher than "*esok*"; it is 9,409 times for "*besok*" and 2,206 times for *"esok*". It is illustrated that the highest number of the usage of the two words "*besok*" or "*esok*" found in 1,152 Bahasa Indonesia clauses in news, 2012. The diagram also describes that the lowest number of using the words "*besok*" and "*esok*" are found in Wikipedia. The following are the examples of the word "*besok*" in sentences.

(16) *Dan tiada seorang pun yang dapat mengetahui (dengan pasti) apa yang akan diusahakannya **besok**.*

(17) *Demikian diungkapkan oleh Michael Isikoff dan David Corn, dua wartawan investigatif dalam buku mereka "Hubris" yang akan beredar **besok**.*

Data (16) and (17) describe that the word "*besok"* means tomorrow. Both sentences describe that the word "*besok*" stands alone at the end of the sentence. It was also informed that the word "*besok*" could be paired with "*hari*" to become "*besok hari*", "*pagi*" to become "*besok pagi*", *"-nya*" to become "*besoknya*", "harinya" to become besok harinya, and "*Senin/ Selasa/ Rabu/ Kamis/ Jumat/ Sabtu/ Minggu*" to become "(the name of the day) *besok"* as described in the following table 3.

"*besok*" (9,409 corpora)					
besok	*besok (hari)*	*besok (pagi)*	*besok (-nya)*	*besok (harinya)*	*(nama hari) besok*
6807	315	587	364	34	1302

Table 3

The words "*besok*" in corpora collection Leipzig University (2008-2018)

(18) *Rencananya* ***besok hari*** *(14/1) pukul 09.00 WIB akan dilakukan proses fit and proper test terhadap Budi Gunawan sebagai calon tunggal Kapo*lri.

(19) ***Besok pagi*** *selama 3 hari Munas Arsada akan berusaha mencari solusi agar tiga aspek utama dalam pelayanan kesehatan, yaitu kualitas pelayanan yang terstandar, akses terhadap pelayanan dan cost yang murah dapat berkombinasi dengan baik*.

(20) *Tanpa menunda,* ***besoknya*** *ia meminta izin tempatnya bekerja, untuk mengurus BPJS*.

(21) Elly bersama Hariska sedang membungkus makanan berupa permen dan makanan `snack` untuk persiapan ulang tahun kesembilan Farhan ***besok harinya*** *(1 Oktober).*

(22) *Saya akan rapat teknis* ***Senin besok*** *dan minta laporan dari daerah untuk HKM ini," ujar Indri*.

While the word "*esok*" could stand alone (725data), could be paired with "*hari*" to become "*esok hari*", "*pagi*" to become "*esok pagi*", "*-nya*" to become "*esoknya*", "*harinya*" to become "*esok harinya*", and as described in the following table 4.

"*esok*" (2,206 corpora)					
esok	*esok (hari)*	*esok (pagi)*	*esok (-nya)*	*esok (harinya)*	*(nama hari) esok*
725	484	191	286	339	181

Table 4

The words "*esok*" in corpora collection Leipzig University (2008-2018)

(23) *Sementara itu Senator Obama dalam kampanye terakhirnya di Virgina mengatakan kepada para pemilih bahwa dia punya satu kata;* ***'Esok***.

(24) *Menurut Djoko,* ***esok hari*** *Kementerian PU akan menggelar rapat dengan para ahli bangunan Jepang itu guna membahas kontruksi bendungan Situ Gintung yang baru*.

(25) ***Esok pagi*** *kami harus menjalani tes kesehatan, sambung Amir*.

(26) Ket*ika Jen ke sekolah dengan gaya rambut baru,* ***esoknya*** *Amy mengganti gaya rambutnya tak mau kalah*.

(27) *Keluarga ini hendak membeli ikan ke Muara Angke, Jakarta Barat, untuk dijual* ***esok harinya***.

(28) *Kemungkinan perbaikan dikerjakan malam hari untuk mengantisipasi penumpang yang masuk kerja lagi* ***Senin esok***.

3.4 The Distribution of the word "*lusa*" in Corpora Collection Leipzig University

The word "*lusa*" based on KBBI is */lu.sa/;* it is a noun. In Bahasa Indonesia means '*hari sesudah besok'* or in English means 'a day after tomorrow'. The usage of the word "*lusa"* in corpora collection was reported for 513 times. The 513 times were found in the several sources, such as blogs, news, newscrawl, web, web public, and Wikipedia. From the six sources obtained, web was reported had the highest number of containing the word "*lusa*"; it is 205 times.

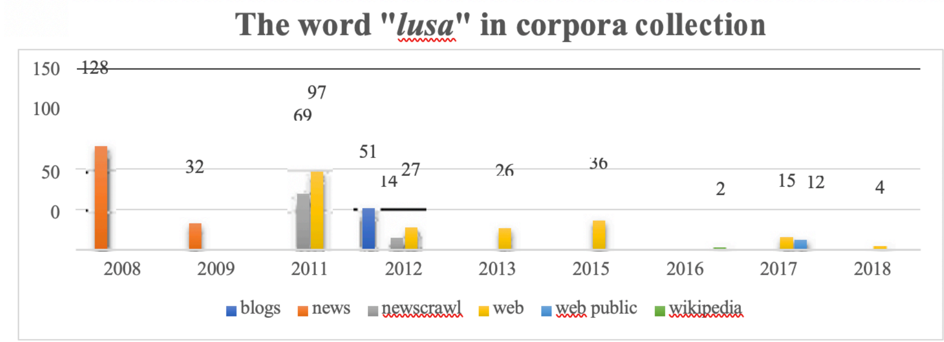


Diagram 4

The words "*lusa*" in Bahasa Indonesia Corpora Collection Leipzig University (2008-2018)

Besides the highest number of using the word "*lusa*", the diagram also describes that Wikipedia has the smallest number of using the word "*lusa*" in Bahasa Indonesia clauses; it is found only two corpora found in 2016. In total, it was reported that the word "lusa" found is 513 data in in corpora collection Leipzig University 2008-2018 (excluded 2014). The following are the examples of the data found.

(29) *Ia mengatakan, tim masih punya waktu sehari untuk recovery (pemulihan tenaga) dan* ***lusa*** *tim akan berangkat ke Bangkok.*

(30) *Selanjutnya orang-orang Jipang itu besok atau* ***lusa*** *harus pergi ke Pajang dengan sebuah pengawalan yang kuat bersama-sama Ki Gede Pemanaha*n.

3.5 The Distribution of the word "*selumbari*", "*tulat*", and "tubin" in Corpora Collection Leipzig University

From the corpora collection, it was reported that there is 7,309,185 corpora in Bahasa Indonesia. From the total number mentioned, there is 72,626 corpora containing the deitic time words in Bahasa Indonesia. The present writer did not find the word "*selumbari*", "*tulat*", and "*tubin*" as one of them.

IV. Conclusion

It was concluded that there are only four Bahasa Indonesia deictic time words ("*kemarin*", "*hari ini*", "*besok/ esok*",and "*lusa*") are found in the data while "*selumbar*i", "*tulat*", and "*tubin*" are not. The deictic time words "*beso*k" and "*esok*" has similar meaning and they could be paired with "*hari*" to become "*besok/ esok hari*", "*pagi*" to become "*besok/ esok pagi*", *-nya* to become "*besoknya/ esoknya*", "*harinya*" to become "*besok/ esok harinya*", and the combination of the days' name to become "*Senin/ Selasa/ Rabu/ Kamis/ Jumat/ Sabtu/ Minggu besok/ esok*".

References

Altohami, W. M. A., & Salama, A. H. Y. (2019). The Journalistic Representations of Saudi Women in the Corpus of Contemporary American English (COCA): A Corpus Critical Discourse Analysis. *International Journal of English Linguistics;*, *9*(6), 320–336.

Baker, P. (1996). Social Involment in corpus studies. In V. Viana, S. Zyngier, & G. Barnbrook (Eds.), *Perspectives on Corpus Linguistics* (pp. 17–28). John Benjamin Publishing Co.

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). Monolingual corpora of standard size. In *Corpus Linguistics 2007* (pp. 1–13).

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications Ltd.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *8th International Language Resources and Evaluation* (pp. 759–765).

Jones, C., & Waller, D. (2015). *Corpus Linguistics for Grammar*. Routledge.

Kwary, D. A. (2019). A corpus platform of Indonesian academic language. *SoftwareX*, *9*, 102–106.

Larasati, S. D., Kubon, V., & Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. In *International Workshop on Systems and Frameworks for Computational Morphology* (pp. 119–129).

Nurlina, W. E. S. (2014). Kosakata Bahasa Jawa sebagai salah satu pengembang kosakata Bahasa Indonesia. *Sawerigading*, *20*(1), 35–43.

Rafatbakhsh, E., & Ahmadi, A. (2019). A thematic corpus-based study of idioms in the Corpus of Contemporary American English. *Asian-Pacific Journal of Second and Foreign Language Education*, *4*(11), 1–21. https://doi.org/10.1186/s40862-019-0076-4

Sujatna, E. T. S., Heriyanto, H., Krisnawati, E., Amalia, R. M., & Pamungkas, K. (2020). Portraying the Word "Tourism" in English: A Corpus Linguistic Study. *Sosiohumaniora*, *22*(2), 181–189.

Sujatna, M. L., Sujatna, E. T. S., & Pamungkas, K. (2019). Exploring the Use of Modal Auxiliary Verbs in Corpus of Contemporary American English (COCA). *Sosiohumaniora*, *21*(2), 166–172.

Tillman, K. A., Marghetis, T., Barner, D., & Srinivasan, M. (2016). Today is tomorrow's yesterday: Children's acquisition of deictic time words. *Cognitive Psychology*, *92*(2017), 87–100.

Wicaksono, A. F., Vania, C., Distiawan, B., & Adriani, M. (2014). Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. In *28th Pacific Asia Conference on Language, Information and Computation* (pp. 185–194).

Yuliawati, S., Dienaputra, R. D., Sujatna, E. T. S., Suryadimulya, A. S., & Lukman, F. (2019). Looking into "Awewe" and "lalaki" in the Sundanese Magazine Mangle: Local Wisdom and a Corpus Analysis of the Linguistic Construction of Gender. *International Journal of Advanced Science and Technology*, *28*(8s), 549–559.

Yusu, X. (2014). On the Application of Corpus of Contemporary American English in Vocabulary Instruction. *International Education Studies*, *7*(8), 68–73.

MANIPURI VERBS: A LEXICOGRAPHICAL CHALLENGE

Heisnam Kenny Devi, Leihaorambam Sarbajit Singh
Department of Humanities and Social Sciences,
Indian Institute of Information Technology Manipur, India
kenny@iiitmanipur.ac.in; sarbajit@iiitmanipur.ac.in

Abstract:

The agglutinating nature of Manipuri affects verb more significantly than other categories. Verbs change morphologically as they receive affixes during extensions or, more precisely, during their inflections. It creates a big challenge for the lexicographer in compiling Manipuri dictionary. Many dictionaries compiled in Manipuri only picture in the adverbial, adjective, nominative form and other applicative as they can be derived from verbs by means of affixation. There is no extension of the meaning as it is claimed that the range of suffixes which can be added to the verb root can be from one to ten (Chelliah, 1997), one to eleven (Devi, 2002). It is due to this reason many dictionaries cannot provide the parts of speech category (POS) of the verb with many suffixes. The usage of the dictionary is to identify the word with its related grammatical relations. It is used as a reference in the proper analysis of the word. It uses qualitative and descriptive method for the study and the dictionaries were analyzed based on the information provided in Manipuri verbs. With this thought it is an attempt to discuss the various issues related to Manipuri verb while compiling dictionaries.

Key words: Agglutinating, Parts of Speech (POS), Tibeto-Burman, verb roots.

1. INTRODUCTION

Manipuri belongs to the Tibeto-Burman language group of Sino-Tibetan language family. It shares some of the features of Tibeto-Burman language such as tone, verb final word order, presence of velar nasal in the initial position of words, postposition in place of preposition, aspect in place of tense etc. The morphology of the language can be broadly studied in two ways: noun morphology (nominal morphology) and verb morphology (verbal morphology). Word formation is mainly contributed by affixation and compounding. It has two distinct lexical categories- noun and verb but other categories are derived from the verb through affixation. Noun morphology mainly deals with the inflectional morphology. Verbal morphology handles both inflectional and derivational morphology. The complexity of verbal morphology lies in the ability of taking both the nominal as well as verbal affixes. However, verbs in Manipuri are not inflected for person, number but inflected for aspect, mood and applicative. The agglutinating nature of Manipuri affects verbs more significantly than other categories as verbs undergo morphological change when they receive affixes. Compilation of Manipuri dictionaries demands a great knowledge of its morphology especially verb morphology.

1.1 EARLIER WORK

The pioneer work in the development of Manipuri Lexicography was done by the British. The first Manipuri involving Dictionary was compiled by George Gordon, the British political Agent of Manipur in the year 1837 (Singh 2011:7). It is a dictionary in English, Hindi and Manipuri. After a century the people started to realize the importance of having Manipuri dictionary and the compilation of Manipuri involving dictionary started to increase in various combinations: English, Hindi and Manipuri. There are more than 40 Manipuri involving dictionaries and which are mostly compiled in Bengali Scripts even though it has its own scripts (called Meetei Mayek or Meetei scripts). The 55 Bengali symbols have been represented by 38 Manipuri phonemes (including two tones) creating lots of phonological problems (Singh and Singh, 2007). Semantic issues are discussed by Singh and Singh (2006) in making English - Manipuri dictionary where English is the source language (SL) and Manipuri is the target language (TL). As "there are no exact correspondences between words in different languages" (Nida, 1958:281), finding the equivalent word is a big issue especially when the two languages are not closely related (linguistically as well as culturally). To sum up, it can be said that most of the Manipuri involving dictionaries compiled so far are meant for the practical aspect neglecting the theoretical aspects. In addition to it, the information provided in the microstructure is not sufficient as compared to the macrostructure provided in the dictionaries.

2. Methods

It is a descriptive in nature. For the purpose of the study, three dictionaries have been chosen

(i) *Manipuri to English Dictionary* by Soibam Imoba in 2004

(ii) *Wahei Kanglon- A Dictionary of Manipuri Verbs* by M.S. Ningomba in 2010 and

(iii) *A Comprehensive Manipuri to English Learner's Dictionary in Meitei Mayek and Bengali Scripts* by Hidam Dolen in 2012). The purpose of the study is to discuss the morpho-syntactic nature of the verb roots entered into the dictionaries.

3. Result

3.1 Description of the Dictionaries

Compiling dictionaries in Manipuri is generally made based on the intuition of the lexicographer focusing less on the intended users of the dictionaries. Considering the three mentioned dictionaries above, the following observations could be obtained as shown in Table 1 below.

Table 1: Table showing the comparison among the three dictionaries.

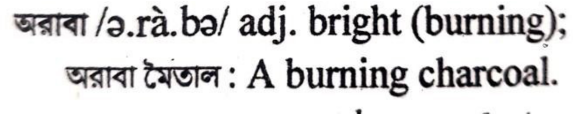
	Abbreviation and Symbols	**Spelling and Transcription**	**Languages used**	**verb root entries and their derivatives**	**Dictionary arrangement**	**Verb extension**
Imoba (2004)	Yes	Yes for both	Manipuri (Bengali scripts) and English	Not mentioned	Bengali alphabetical sorting order	No.
Ningomba (2010)	Yes	Yes for spelling, Transcription doesn't follow IPA pattern	Manipuri (Meetei Mayek) and English	More than one thousand	4 vowels and 15 consonants in the order of i, u, o, aw, p, t, k, ph, th, kh, m, n,ng,ch,s,l,h,w and j.	No.
Dolen (2012)	Yes	Yes for both	Manipuri (Meetei Mayek and Bengali scripts) and English	Not mentioned	Meetei Mayek alphabetical sorting order	No

It can be claimed that these dictionaries are meant for communication-oriented purpose as compared to knowledge-oriented one. But the morpho-syntactic properties of the language are not seen.

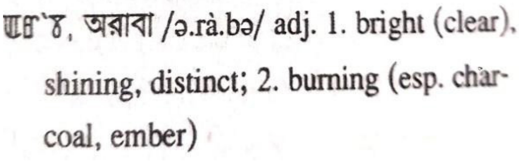
4. Finding and Discussion

All these dictionaries should not be mixed in finding a word as they are different in their arrangement. In the first dictionary, Imoba (2004), arrangement follows the pattern of Bangla / Assamese scripts. Every verb root entry is followed by its adverb and verbal noun but the extension of meaning is not clearly visible. As claimed by Singh (2000: 35), Manipuri has 25 verbal suffixes which play significant role in the word formation process. It is viable to create a large corpus for Manipuri, if these suffixes added to the verb root. It is important for the lexicographer to locate a word with its grammatical information (especially morpho-syntactic) to the user. If a user wants to determine the parts of speech (POS) of a word which is made by the combination of multi suffixes to the verb root, their derivative information should be provided in the dictionary. For instance the word such as *jeŋ.niŋ.həl.lu.bə.də.gi.ni 'for the sake of causing someone to look into the matter'* is formed by the addition of 6 suffixes to the verbal root *jeŋ* (see) as *jeŋ* (VR) (see) + *niŋ* (MD) + *həl* (CAU) + *lu* (IMP) + *bə* (NOM) + *də.gi* (ABL) + *ni* (COP). One of the expected reasons worth mentioning is that affix ordering is flexible in some condition and in some case it obeys some sort of hierarchy. . For instance, mood should always precede habitual aspect that is *cá* (eat) + *niŋ* (MD) + *gəl* (HAB) + *li* (ASP) > *cá.niŋ.gəl.li* (one who is in the mood of habitual eating). The lexicographers don't stretch their knowledge and pay less attention to such areas. Other prominent feature observed from these dictionaries is that phonological assimilation is not clearly visible. For instance, all the verb roots begin with lateral /l/ changes into /ɹ/ when the attributive prefix /ə/ is added to them in their adjectival form. That's why *lan.bə* 'wrong', *lau.bə* 'loud' become *ə.ɹan.bə* and *ə.rau.bə*. But it would be wrong to consider *ɹan* and *ɹau* as verbal root. In Ningomba (2010), the attributive prefix /ə/ is not found which is considered to be productive even though the dictionary is specially meant for verbs.

The entry word /ə.ɹa.bə/ 'bright' is highlighted in Imoba (2004: 21) and Dolen (2012: 687) in the following manner



অরাবা /ə.rà.bə/ adj. bright (burning);
অরাবা মৈতাল : A burning charcoal.



ꯑꯔꯥꯕ, অরাবা /ə.rà.bə/ adj. 1. bright (clear), shining, distinct; 2. burning (esp. charcoal, ember)

It seems like the only difference between the two dictionaries is the addition of Meetei Mayek (Meetei Script) in the second dictionary. Moreover, grammatical information related to transitivity and intransitivity of the verb root is not highlighted for the homophonous verb. Zgusta (1971: 15) mentions that the 'the theory of lexicography is connected with all the disciplines which study the lexical system: semantics, lexicology, grammar, stylistics'. Manipuri verb dictionaries try to picture only semantics and its related affairs. 'The first and the basic purpose of indicating grammatical information in the dictionary is to indicate the morpho- syntactic peculiarities of the lexical unit'(Singh,1991:116). To provide grammatical information of the head entry especially for the language like Manipuri is a big challenge to be tackled by the lexicographer in compiling a dictionary. Verbs also can be transitive and intransitive. Transitive verbs take objects and intransitive verbs do not. Words in Manipuri dictionaries provide only verbs whether it is transitive or intransitive is not shown. Most Manipuri involving dictionaries do not provide detail grammatical information of the entry. In a Learner's dictionary there should be more grammatical information than in general purpose dictionary. However, special dictionaries like pronouncing and orthographical dictionaries do not require grammatical information.

5. Conclusion

In this paper we have shown that Manipuri dictionaries cannot depict verbal extension despite of so many verbal suffixes. The sources of verb root description and their inflected and derivative forms are not adequately explained. A real-compiled Manipuri verb dictionary, especially for learners must show this state of affairs. In order to achieve this goal, the lexicographers need to enter all the possible information which not only helps in enhancing the corpus, but also in further language technology development applications.

Symbols and Abbreviation

ABL	-	Ablative
ASP	-	Aspect
CAU	-	Causative
COP	-	Copula
HAB	-	Habitual
IMP	-	Imperative
MD	-	Mood
NOM	-	Nominative
VR	-	Verbal root

REFERENCES

Chelliah, Shobhana, L. (1997). A *Grammar of Meithei*. New York: Mouton de Gruyter, Berlin. Devi, M. Bidyarani. (2002). *Manipuri verbs*. Unpublished dissertation. Manipur University.

Nida, E. A. (1958). Analysis of Meaning and Dictionary Making. In *International Journal of American Linguistics*. vol. 24

Singh, Ch. Yashawant. (2004). *Manipuri Grammar*. New Delhi: Rajesh Publication.

Singh, L. S., & Singh, S. I. (2006). IN MAKING ENGLISH-MANIPURI DICTIONARY--THE SEMANTIC PROBLEMS. *Language in India*, *6*(10).

Singh, L. S., & Singh, S. I. (2007). PHONOLOGICAL PROBLEMS IN MAKING ENGLISH- MANIPURI DICTIONARY FOR MANIPURI SPEAKERS. *Language in India*, *7*(9).

Singh, L. Sarbajit. (2011). *Dictionary-Making: English-Manipuri.* Akansa New Dehli: Akansha Publishing House.

Singh, Ram. Adhar. (1991). *An Introduction to Lexicography*. Central Institute of Indian languages, Mysore.

Zgusta, L. **(**1971). *Manual of Lexicography.* The Hague/Paris: Mouton.

Dictionary Cited:

Soibam, Imoba. (2004). *Manipuri to English Dictionary*. Imphal.

Ningomba, M.S. (2010). *Wahei Kanglon- A Dictionary of Manipuri verbs*. Imphal.

Hidam, Dolen. (2012). *A Comprehensive Manipuri to English Learner's Dictionary in Meitei Mayek and Bengali Scripts*. Imphal.

APPENDIX I

Meetei Mayek 27 Letters

ꯏꯌꯦꯛ ꯏꯄꯤ /ijek ipi/ **Main Letters**

ꯀ	ꯀꯣꯛ	ꯁ	ꯁꯝ	ꯂ	ꯂꯥꯏ
/k/	/kok/	/s/	/səm/	/l/	/lai/
ꯃ	ꯃꯤꯠ	ꯄ	ꯄꯥ	ꯅ	ꯅꯥ
/m/	/mit/	/p/	/pa/	/n/	/na/
ꯆ	ꯆꯤꯜ	ꯇ	ꯇꯤꯜ	ꯈ	ꯈꯧ
/c/	/cil/	/t/	/til/	/kʰ/	/kʰəu/
ꯉ	ꯉꯧ	ꯊ	ꯊꯧ	ꯋ	ꯋꯥꯏ
/ŋ/	/ŋəu/	/tʰ/	/tʰəu/	/w/	/wai/
ꯌ	ꯌꯥꯡ	ꯍ	ꯍꯨꯛ	ꯎ	ꯎꯟ
/j/	/jaŋ/	/h/	/huk/	/u/	/un/
ꯏ	ꯏ	ꯐ	ꯐꯝ	ꯑ	ꯑꯇꯤꯌꯥ
/i/	/i/	/pʰ/	/pʰəm/	/ə/	/ə.ti.ja/
ꯒ	ꯒꯣꯛ	ꯓ	ꯓꯝ	ꯔ	ꯔꯥꯏ
/gok/	/gok/	/ɟʰ/	/ɟʰəm/	/ɹ/	/ɹai/
ꯕ	ꯕꯥ	ꯖ	ꯖꯤꯜ	ꯗ	ꯗꯤꯜ
/b/	/ba/	/ɟ/	/ɟil/	/d/	/dil/
ꯘ	ꯘꯧ	ꯙ	ꯙꯧ	ꯚ	ꯚꯝ
/gʰ/	/gʰəu/	/dʰ/	/dʰəu/	/bʰ/	/bʰəm/

APPENDIX II

ꯂꯣꯟꯁꯨꯝ ꯏꯌꯦꯛ /lonsum ijek/ Final consonants

ꯛ	ꯀꯣꯛ ꯂꯣꯟꯁꯨꯝ	ꯜ	ꯂꯥꯏ ꯂꯣꯟꯁꯨꯝ	ꯝ	ꯃꯤꯠ ꯂꯣꯟꯁꯨꯝ
/-m/	/kok lon.sum/	/-l/	/lai lonsum/	/-m/	/mit lonsum/
ꯞ	ꯄꯥ ꯂꯣꯟꯁꯨꯝ	ꯟ	ꯅꯥ ꯂꯣꯟꯁꯨꯝ	ꯠ	ꯇꯤꯜ ꯂꯣꯟꯁꯨꯝ
/-p/	/pa lonsum/	/-n/	/na lonsum/	/-t/	/til lonsum/
ꯡ	ꯉꯧ ꯂꯣꯟꯁꯨꯝ	ꯢ	ꯏ ꯂꯣꯟꯁꯨꯝ		
/-ŋ/	/ŋəu lonsum/	/-i/	/i lonsum/		

APPENDIX III

ꯆꯩꯇꯥꯞ ꯏꯌꯦꯛ /cəitəp ijek/ Dependent vowel signs

ꯣ /o/	ꯑꯣꯅꯥꯞ /o.nəp/	ꯤ /i/	ꯏꯅꯥꯞ /i.nəp/	/a/	ꯑꯥꯅꯥꯞ /a.nəp/	ꯦ /e/	ꯌꯦꯅꯥꯞ /je.nəp/
ꯧ /əi/	ꯁꯧꯅꯥꯞ /səu.nəp/	/u/	ꯎꯅꯥꯞ /u.nəp/	/əi/	ꯆꯩꯅꯥꯞ /cəi.nəp/	/ŋ/	ꯅꯨꯡ /nuŋ/

ꯈꯨꯗꯝ ꯏꯌꯦꯛ /kʰudəm ijek/ Punctuation

꯫ Full stop	ꯆꯩꯈꯩ /cəikʰəi/	꯬ Heavy tone	ꯂꯨꯝ ꯏꯌꯦꯛ /lum ijek/	killer	ꯑꯄꯨꯟ ꯏꯌꯦꯛ /əpun ijek/

[illegible] [illegible] /cəisiŋ ijek/ **Digits**

𖫀 Zero	[illegible] /pʰun/
𖫁 One	[illegible] /əmə/
𖫂 Two	[illegible] /əni/
𖫃 Three	[illegible] /əhum/
𖫄 Four	[illegible] /məɹi/
𖫅 Five	[illegible] /məŋa/
𖫆 Six	[illegible] /təɹuk/
𖫇 Seven	[illegible] /təɹet/
𖫈 Eight	[illegible] /nipal/
𖫉 Nine	[illegible] /mapəl/

Sorting order

Main Letters

[illegible]

Final consonants

[illegible]

Dependent vowel signs [illegible]

[illegible]

ON THE INCLUSION OF EPONYMS IN ONLINE ENGLISH LEARNER'S DICTIONARIES

Jinhong Huang
Fudan University, China
jhhuang19@fudan.edu.cn

Abstract

Eponymy is a word-formation process where a novel word is derived from proper names. Eponyms represent an interesting, yet not sufficiently studied linguistic phenomenon. English dictionaries include some established eponyms. *Oxford Advanced Learner' s Dictionary*, *Macmillan English Dictionary for Advanced Learners*, and *Cambridge Advanced Learner's Dictionary* are three representative advanced English learner's dictionaries and all of them have launched their online versions. This paper, based on the collection, comparison and analysis of eponyms in the three dictionaries, finds that the eponyms in the dictionaries are largely based on word-building and their ways of formation are diversified. Most eponyms in the said dictionaries are based on real names or personal names from other countries, while some others are built on fictional names or anthroponyms from English-speaking countries. The dictionaries leave room for improvement as regards the treatment of eponyms. Namely, there is the unsystematic inclusion of eponyms, the inadequacy of including relatively new eponyms and the inconsistency of offering etymological information for eponyms. This paper proposes suggestions for betterment. The three dictionaries should enhance systematic coverage of eponyms, better the inclusion of relatively new eponyms and improve the consistency of furnishing etymological information. The research on eponyms in the online English learner's dictionaries can contribute to the betterment of eponym treatment and can facilitate our perception of the word-formation per se from the lexicographic perspective.

Keywords: eponym, online English learner's dictionaries, inclusion of headwords

1 Introduction

Eponymy is the process in which words for activities, inventions, places and so on have been derived from personal names. Eponyms (epi- 'upon', onym 'name') belong to a particular and ever-expanding lexical category. According to the online version of *The Oxford English Dictionary* (henceforth *OED*), the term "eponym" cropped up as early as the year 1885. Eponyms have continually contributed to English vocabulary and they currently constitute a considerable proportion of the English word stock.

Some articles and monographs on word-formation or lexicology have touched upon English eponyms (e.g. Štekauer 1997; Lalić 2004; Gao 2012). They focus on etymologies, formation, or rules of the words in question. "Eponym" was not incorporated in the title of the first dictionary concerning the stories of some words derived from names of literary heroes (Edwards 1968). Hendrickson (1972), after four years, authored the first lexicographic work to bear this term. There was a debate on whether proper names deserved to be included in the dictionary as much as common nouns (e.g., Landau 1984; Mufwene 1988; Urdang 1996). However, proper names and eponyms originating from them, undoubtedly, have been partly welcomed into the dictionary proper. Former studies into eponyms, in most cases, were conducted by gathering eponyms from various resources or enumerating typical instances. Nevertheless, research on the eponyms included in English dictionaries is rare.

Oxford Advanced Learner's Dictionary, *Longman Dictionary of Contemporary English*,*Macmillan English Dictionary for Advanced Learners*, *Collins COBUILD English Language*

Dictionary and *Cambridge Advanced Learner's Dictionary*, known as "Big Five", are mainstream large-scale monolingual English learner's dictionaries (henceforth *MELDs*) aiming at foreign English learners and all of them have launched their online versions (henceforth referring to *OALD*, *LDOCE*, *MEDAL*, *CCELD*, *CALD* as the five online dictionaries).

Eponyms have been applied to many fields of knowledge, for example, medicine, history, geography and so forth. Such words largely derive from the names of real people, but some from literary or mythical heroes. In this sense, eponymy is a way of preserving the cultural heritage of a nation and a society, even the cross-cultural legacy. Since the cultural specificity of most eponyms partly results in the semantic opacity of this lexical category, eponyms pose a challenge to the acquisition of English learners, especially non-natives. Therefore, the current study will zoom in on the eponyms in some *MELDs* and examine how the dictionaries treat the words in question, which hopefully, can better the coverage and presentation of the words in the said dictionaries. Furthermore, the study may also help to paint a picture of eponyms and the word-formation itself from the angle of lexicographic products.

2 Eponyms and online English learner's dictionaries

From the outset, it is necessary to examine the different views on the definition of eponymy and the types of formations considered as eponyms. A chronological trace of the defining of eponyms finds no full agreement among researchers, whose definitions, however, can be summarized into two categories.

The first kind is featured by the broader definition. For example, the definition of an eponym in the *OED* is "proper name used generically; more loosely, the generic name itself, or any noun phrase of specific meaning which includes a proper name", which may include anthroponyms, toponyms and trademarks. This definition is endorsed by researchers like Fromkin et al. (2003) and Minkova & Stockwell (2009). Fromkin et al. (2003: 95) took the view that "eponyms are words derived from proper names and are another of the many creative ways that the vocabulary of a language expands". The latter, albeit providing no exact definition of eponyms, classified such words into those "based on personal names", "based on geographical names", "based on names from literature, folklore, and mythology" and "based on commercial brand names " (Minkova & Stockwell 2009: 19-21).

The second type centers on narrower definitions and those of McArthur (1996) and Crystal (2008) are cases in point. According to McArthur (1996), eponym was used to refer to "a personal name from which a word has been derived, the person whose name is so used and the word so derived" (ibid: 350). In this context, toponyms and trademarks are excluded.

Table 1 The definitions of eponyms in three learner's dictionaries

Dictionary	Definition
CALD	the name of an object or activity that is also the name of the person who first produced the object or did the activity
MALD	a person that a place, discovery, era or invention is named after, for example 'Elizabethan' or 'braille'; a product name that is frequently used to replace a particular item, such as 'aspirin' for painkiller or 'hoover' for vacuum cleaner (This meaning is based on one submitted to the Open Dictionary[1] by: Kerry Maxwell from United Kingdom on 18/04/2018)
OALD	a person or thing, or the name of a person or thing, from which a place, an invention, a discovery, etc. gets its name

[1] The open dictionary in the *MEDAL* is where the public can suggest neologisms.

Among the five *MELDs*, merely the *CALD*, the *MEDAL* and the *OALD* cover the word “eponym” and offer it a definition, as shown in Table 1. The definitions of eponyms in the *MALD* and the *OALD* include the names of products or things as sources in addition to personal names. Lexemes treated as eponyms in this paper are mainly those formed through names of people, real or fictitious.

The *CALD*, the *MEDAL* and the *OALD* are chosen as the objectives of present research to perform case studies into eponyms in *MELDs*. As an attempt at delving into eponyms from the lexicographic perspective, the goals of this paper are twofold:

(1) To classify the eponyms in the *CALD*, the *MEDAL* and the *OALD*;

(2) To identify the problematic treatment of eponyms in the three dictionaries and to put forward measures for improvement.

Accordingly, this paper will primarily deal with three research questions:

(1) What are the eponyms included in the *CALD*, the *MEDAL* and the *OALD* and their categorization?

(2) What problems can be identified regarding the treatment of eponyms in the three dictionaries?

(3) How can we improve the treatment of eponyms?

3 Methods

Three methods have been employed to address the above research questions. First, the present author conducted a quantitative collection of all the eponyms in the *CALD*, the *MEDAL* and the *OALD*. Second, the author adopted a classificatory approach to probing into the eponyms in the three dictionaries, through which we could further compare the inclusion of different types of eponyms in the dictionaries. And the problems in eponym coverage and presentation were analyzed qualitatively.

The data accumulation was based on the definition of the eponym adopted by this paper and the etymological information provided by the *MEDAL* and the *OALD* (the *CALD* does not furnish such information), as well as other four online large-scale dictionaries, i.e. *Collins English Dictionary* (henceforth *CED*), *Merriam-Webster's Collegiate Dictionary* (henceforth *MWCD*), *Oxford Dictionary of English* (henceforth *ODE*) and the *OED*. The four dictionaries were selected as they, in most cases, offer detailed etymological information for headwords. Hence, they can provide reference for the determination of eponyms.

The present author has established a database consisting of 539 eponyms from the *CALD*, 416 from the *MEDAL* and 437 from the *OALD*[2]. If we estimate that the headword coverage in the three dictionaries is 60,000 on average, then it can be counted that eponyms occupy approximately 0.7%- 0.9% out of the total inclusion.

4 Eponyms in the three online English learner's dictionaries

4.1 A classification of eponyms based on their formation

The eponyms in the *CALD*, the *MEDAL* and the *OALD* are either directly derived from personal names and take the form of simple words or are formed through word-formation processes.

[2] Eponymous units at the level of phrases and idioms are not considered in the present study.

Table 2 A classification of eponyms based on their formation

Ways of formation		*CALD*	*MEDAL*	*OALD*
Simple word	number	163	122	128
	proportion	30.24%	29.33%	29.29%
Compounding	number	221	175	183
	proportion	41.00%	42.07%	41.69%
Derivation	number	140	102	109
	proportion	25.97%	24.52%	24.83%
Blending	number	8	5	6
	proportion	1.48%	1.20%	1.37%
Clipping	number	3	8	8
	proportion	0.56%	1.92%	1.82%
Conversion	number	4	4	3
	proportion	0.74%	0.96%	0.68%

(1) Eponyms in the form of simple words

Simple eponyms, a group of words undergoing almost no morphological changes, comprise approximately 30% of this word category in the three dictionaries. They originate from personal names or nicknames. The initial letter of the words is either capitalized or lower-case. For example, "coulomb" was named after the French physicist Charles-Augustin de Coulomb and its first letter is in the lower case. By contrast, "Fahrenheit", named after the Prussian scientist Daniel Gabriel Fahrenheit, starts with a capitalized "F" and makes no modification to the surname of the scientist.

(2) Eponyms from compounding

Eponyms created by compounding make up the largest proportion (above 41%) among all the eponyms from word-building in the three dictionaries. Compound eponyms can be divided into five categories from the perspective of morphology, among which the most common form is written separately. Typical compounds are composed of a proper name and a common word, such as Achilles heel, Alexander technology, Bailey bridge, Copernican system, Doberman pincher and Electra complex. The second type is characterized by juxtaposing two elements without space between them, like Bluetooth (Blue + tooth), loganberry (logan + berry), Frankenfood (Franken- + food) and Corbynomics (Corbyn + -nomics). The sources of the former two eponyms are words, while the source elements of "Frankenfood" are a combining form derived from a proper name (Frankenstein) as well as a word. In contrast, Corbynomics consists of an anthroponym (Jeremy Corbyn) and a combing form originating from economics.

Compound eponyms in the dictionaries are also connected by a hyphen, for instance, Arnold-Chiari malformation, eve-teasing and Sapir-Whorf hypothesis. Amongst all the compound eponyms, those with a possessive case are dominant. Most of the eponyms concern some disciplines, especially medicine. These eponyms are often made up of the names of doctors who found the disease and the words "syndrome, disease or tumor", such as Buerger's disease, Edwards' syndrome, and Huntington's chorea. There are also compound eponyms with a hyphen and a possessive case. For example, Herzberg's two-factor theory and Non-Hodgkin's lymphoma.

(3) Eponyms from derivation

Next to compounding, derivation is the most productive process (about 25% on average) among all the eponyms created by word-formation in the three dictionaries. Eponyms of this kind may be realized by dropping parts of original names and adding suffixes to the names. In some cases, those of Latin and Greek origins drop their endings, as manifested by the examples of Aeolian, Confucian, Sapphic in Table 3. The most frequent suffixes to form eponyms in the learner's dictionaries are - ian, -ish, -ism, -ist and -ize.

Table 3 Examples of eponyms formed through suffixation

Suffix	Example
-ian	Aeolian (Aeolus + -ian), Bayesian (Bayes + -ian), Confucian (Confucius + -ian), Dickensian (Dickens + -ian)
-ism	Maoism (Mao + -ism), Marxism (Marx + -ism), Spoonerism (Spooner + -ism), Thatcherism (Thatcher + -ism)
-ist	Buddhist (Buddha + -ist), Leninist (Lenin + -ist), Maoist (Mao + -ist)
-ite	Bakelite (Baekeland + -ite), Luddite (Ludd + -ite), Thatcherite (Thatcher + -ite)
-ic	Cyrillic (Cyril + -ic), mesmeric (Mesmer + -ic), Platonic (Plato + -n- + -ic), Sapphic (Sappho + -ic)

(4) Eponyms from blending

It is not uncommon in English that an eponym can be formed through the blending of a proper name and a common word. However, such words are often in nonce forms and few of them become permanent English vocabulary. The eponyms of this sort are in a minority in the three dictionaries, such as gerrymander (Gerry [Elbridge Gerry of Massachusetts] + salamander), Linux (Linus [Linus Torvalds] + Unix), Magpie (Mag [Margaret] + archaic pie) and Reaganomics (Reagan [Ronald Reagan] + economics).

(5) Eponyms from clipping

Although clipping does not operate frequently on personal names, several examples have been found among the eponyms in the three dictionaries, like Alzheimer's [< Alzheimer's disease], Down's [< Down's syndrome], Lord's [< Lord's Cricket Ground], and Parkinson's [< Parkinson's disease]. They are mainly possessives, clipping the nouns of their base words and retaining the anthroponyms.

(6) Eponyms from conversion

Conversion (also known as zero derivation) is "a word-building process which involves shifting a word from a word class into another word class, or to put it differently, changing the lexical category of a word" (Bejan 2017: 62). Eponyms of this category only take up a minor percentage (all below 1%) in the three dictionaries. Typical instances are bogart (verb, from the actor Humphrey Bogart), boycott (verb, from the name of Captain Charles C. Boycott, a land agent in Ireland), silhouette (verb, from the name of

Étienne de Silhouette, a French author and politician). The eponyms are formed by converting the nominal personal names into verbs.

4.2 A classification of eponyms based on their etymologies

The eponyms in the *CALD*, the *MEDAL* and the *OALD* can be categorized according to the characteristics of their source names as well.

4.2.1 Eponyms based on real or fabricated names

(1) Eponyms derived from real names

Names of real people are the most common sources of eponymous lexemes, constituting a percentage of over 80% or nearly 90% in the three dictionaries. Anthroponyms or their derivatives have conspicuous cultural associations. Eponymic units based on anthroponyms of celebrities from different fields represent the largest group. Take "Celsius", as an example, it is named after the Swedish astronomer Anders Celsius (1701-1744), who first proposed the centigrade scale in 1742. Gallup poll is another prototypical case, which is derived from George H. Gallup (1901-1984), the American statistician who devised the method.

(2) Eponyms derived from fictional names

The three dictionaries also cover some eponyms (around 13%-17%) built on character names from mythology, folklore, literature, movies and television products. Some onomastic studies fix upon the names of personified images from ancient Greek and Roman mythology, personages of historical legends, which have turned into common names and enriched modern English vocabulary. A small group of eponyms fall into medical and psychological terms, like Achilles tendon, Electra complex, Oedipus complex, narcissism and so on. Uncle Tom means "a black person who is considered to be too eager to agree with white people or too willing to be treated in a way that is not equal to white people" and it is originally the hero in the novel *Uncle Tom's Cabin*. Daisy dukes, now referring to "very short trousers that end just below the hip, made from jeans that have been cut short", comes from the apparel of Daisy Duke, a fictional character in the television series *The Dukes of Hazzard*.

Table 4 A classification of eponyms based on their etymologies

Dictionary	**Real name**		**Fictitious name**	
	Number	**Proportion**	**Number**	**Proportion**
CALD	453	84.04%	86	15.96%
MEDAL	344	82.69%	72	17.31%
OALD	379	86.73%	58	13.27%

4.2.2 Eponyms based on personal names from English-speaking countries or others

Eponyms can also be grouped according to the places where the personal names are from. About 33-38% of the eponyms in the three dictionaries are built on characters or personages from English-speaking countries. For example, "joule" is named after English physicist James Prescott Joule (1818-1889) and "newton" is derived from the English scientist Sir Isaac Newton. In comparison, most eponyms (62%-67%) are based on personal names from other countries. A good case in point can be “pasteurization”. It is from French pasteurization that is based on the surname of its inventor, Louis Pasteur. Some borrowed eponyms even stem from personal names from different countries, like Arnold-Chiari malformation (named after Austrian pathologist Hans Chiari and German pathologist Julius Arnold).

Table 5 A classification of eponyms based on their etymologies

Dictionary	**English Eponym**		**Borrowed Eponym**	
	Number	**Proportion**	**Number**	**Proportion**
CALD	204	37.85%	335	62.15%
MEDAL	148	35.58%	268	64.42%
OALD	145	33.18%	292	66.82%

From the above analyses, it can be summarized that the eponyms in the *CALD*, the *MEDAL* and the *OALD* are largely based on word-formation and the ways of formation are diversified. Most eponyms in the three dictionaries are built on real personal names or those from other countries, while some others are grounded in fictional names or anthroponyms from English-speaking countries.

5 The problematic treatment of eponyms in the three dictionaries

Upon perusal of eponyms in the *CALD*, the *MEDAL* and the *OALD*, the present author has identified three problems regarding the inclusion and presentation of eponyms as follows.

5.1 The unsystematic inclusion of eponyms

The unsystematic inclusion of headwords is a perennial problem in dictionaries and the three learner’s dictionaries are not immune from this. Dictionary-makers omit some co-hyponymic eponyms. A selection of eponyms in the co-hyponymic category finds the exclusion of some of them in the *CALD*, the *MEDAL* and the *OALD*. These words all have frequency of occurrence in the corpus of News on the Web (henceforth *NOW*) and have entered the wordlist of at least one of the said learner’s dictionaries. However, it is found they may not be included in others. For example, among the four eponyms pertinent to “disease”, only the *CALD* includes all of them. “Chagas disease” is absent from the *MEDAL* and “Kawasaki disease” remains on the site of the open dictionary. The *OALD* merely covers “Hodgkin’s disease”, whereas the other three eponymic units are not present in the dictionary. There exists the unsystematic inclusion of eponyms related to “scale” as well. “Richter scale” and “Scoville scale” have been part of the A-Z list of the *CALD*, but “Kelvin scale” hasn’t gained the same status. And this eponym is also not found in the *MEDAL*. “Scoville scale”, like “Kawasaki disease”, is under the pending decision in the open dictionary. As for the *OALD*, we cannot consult the word “Scoville scale” therein.

Table 6 Examples of co-hyponymic eponyms

Dictionary **Eponym**	**Frequency in NOW**	***CALD***	***MEDAL***	***OALD***
Huntington’ s disease	176	√	√	×
Chagas disease	672	√	×	×
Kawasaki disease	1941	√	OPEN DICTIONARY	×
Hodgkin’ s disease	16	√	√	√

(the) Beaufort scale	134	√	√	√
Kelvin scale	39	×	×	√
(the) Richter scale	4154	√	√	√
Scoville scale	204	√	OPEN DICTIONARY	×

5.2 The inadequacy of including relatively new eponyms

Based on eight randomly selected eponyms that were created in the 20th or the 21st century, it is found that for the neologisms at least included in one of the three learner’s dictionaries may not be covered in the others. For example, “Bechdel test”, coined in 2007 according to the *OED*, has the frequency of 792 times in the *NOW* corpus. However, it is not included in the *OALD*. Similarly, “burpee” is absent from the *MEDAL*. Three co-hyponymic neological eponyms can merely be consulted in the *MEDAL*. It is worth mentioning that “Obamacare” and “Trumpism”, two eponymous lexemes relevant to politicians, both have cropped up in the 21st century and have high frequency in the *NOW* corpus. But they cannot be looked up in the *OALD*.

Table 7 Examples of new eponyms

Dictionary **Eponym**	**Date of first appearance**	**Frequency in NOW**	***CALD***	***MEDAL***	***OALD***
Bechdel test	2007	792	√	√	×
burpee	1939	1029	√	×	√
copernicium	2009	58	×	√	×
flerovium	2012	79	×	√	×
livermorium	2012	38	×	√	×
mondegreen	1954	75	×	√	×
Obamacare	Early 21st century	50915	√	OPEN DICTIONARY	×
Trumpism	2015	6195	√	OPEN DICTIONARY	×

5.3 The inconsistency of offering etymological information for eponyms

The online dictionary is generally not limited by physical size to provide information. The *MEDAL* and the *OALD* both offer etymological information for some eponyms in the columns of “CULTURAL NOTE” and “Word Origin” respectively. However, the *CALD* does not provide information of this kind. Such formative information is not always consistently furnished by the *MEDAL* and the *OALD*. As displayed in Table 8, the *MEDAL* merely provides a cultural note for “Achilles heel” and the word origin of “Asperger’s syndrome” is not supplied in the *OALD*.

Table 8 Etymological information for eponyms in the dictionaries

Dictionary **Eponym**	***CALD***	***MEDAL***	***OALD***
Achilles (’) heel	/	CULTURAL NOTE When the mythical Greek hero Achilles was a baby his mother dipped him in the River Styx to protect him from all injuries. But she held him by one heel which therefore was not protected. He was eventually killed by an arrow that hit his heel.	Word Origin Named after the Greek hero Achilles. When he was a small child, his mother held him below the surface of the river Styx to protect him against any injury. She held him by his heel, which therefore was not touched by the water. Achilles died after being wounded by an arrow in the heel.
Adam’s apple	/	/	mid 18th cent.: so named from the notion that a piece of the forbidden fruit became lodged in Adam’s throat.
Asperger’s syndrome	/	/	/
Don Juan	/	/	From the name of a character from Spanish legend who was skilled at persuading women to have sex with him.
hertz	/	/	late 19th cent.: named after H. R. Hertz (1857-94), the German physicist and pioneer of radio communication.
leotard	/	/	early 20th cent.: named after Jules Léotard (1839–70), French trapeze artist.

6 Suggestions for better treatment

Given the problematic treatment of eponyms in the *CALD*, the *MEDAL* and the *OALD*, this paper proposes some suggestions for betterment.

6.1 To enhance the systematicness of eponym coverage

The vocabulary of a language is systematic, and the wordlist in a dictionary can also be taken as a systematic whole. Co-hyponymic eponyms are better to be systematically included to construct a desirable macrostructure for the said dictionaries as well as increase the search possibilities of the words in question. Hence, for the co-hyponymic eponyms, such as those concerned with “disease” and “scale” listed in Table 6 covered by at least one dictionary among the three, they are supposed to be considered by the other two dictionaries.

6.2 To better the inclusion of relatively new eponyms

The *CALD*, the *MEDAL* and the *OALD* are not confined by the publishing cycle, compared to printed dictionaries. They can quickly add new words and meanings. Lexicalization, according to Bussmann (1996: 681), means "the adoption of a word into the lexicon of a language as a usual formation that is stored in the lexicon and can be recalled from there for use". From a similar historical perspective, Brinton and Traugott (2005: 45) held that institutionalization denoted the dissemination of a usage to a community and its establishment as the norm. Lexicalization has more to do with word structure, whereas institutionalization emphasizes the socio-pragmatic side of words. Some eponyms once in nonce forms will undergo lexicalization and institutionalization and then become part of the permanent vocabulary supposing that they continue to be used in the language. Such eponyms, partially, have been incorporated into general dictionaries, including *MELDs*.

Some relatively new eponyms, if supported by ample evidence of lexicalization and institutionalization, should be considered in terms of their inclusion. In this case, the dictionaries can reflect the dynamic development of words in a timelier manner. Furthermore, the availability of new eponyms can be expanded to satisfy the user needs, so as to improve the usability of dictionaries.

6.3 To improve the consistency of providing etymological information

Online dictionaries, an integral part of electronic dictionaries, are relatively free from the limitation of space to provide information. As early as in the 1990s, Hacken (1998: 16) suggested "what is required in electronic dictionaries is an explicit treatment of word formation. It would be highly desirable for such a treatment to fit in with the concept of reusability to enhance thereby the consistency of the dictionary". He (2006: 254) further pointed out that "the adequate representation of word formation in a learners' dictionary is an important asset in the acquisition of vocabulary." For a user who does not know about Achilles, Adam or Asperger, the notes on the word origin concerning brief introductions to these characters should be a helpful starting point beyond which he/she may want to check in an encyclopedia or other online sources for more information to help him/her better understand the relevant entries Achilles (') heel, Adam's apple, Asperger's syndrome and so forth. The *CALD*, the *MEDAL* and the *OALD*, accordingly, should improve the etymological information to serve the pedagogical and reference purpose of the learner's dictionaries, thereby facilitating the comprehension of the specific words.

7. Concluding remarks

Eponymy is an interesting way of word-formation. New eponyms are being formed all the time since there appear to be no limitations on the productivity of eponymous formations. Dictionaries such as the online English learner's dictionaries have included some lexicalized and institutionalized eponyms. As eponyms emerge against a well-defined cultural background, they involve rich cultural information. Although eponyms do not account for a large part of vocabulary in the said dictionaries, their appropriate treatment that can contribute to the linguistic and cultural knowledge acquisition of non-native English learners should not be ignored. Online English learner's dictionaries are supposed to give full play to its advantages in the capacity of headword coverage and treat eponyms consistently and systematically, thereby increasing their data availability and usability.

References

A. Dictionaries

CALD (Cambridge Advanced Learner's Dictionary). Accessed on 20 May 2021.https://dictionary.cambridge.org/

CCELD (Collins COBUILD English Language Dictionary). Accessed on 20 May 2021.https:// www.collinsdictionary.com/

CED (Collins English Dictionary). Accessed on 20 May 2021. https://www.collinsdictionary.com/. LDOCE (Longman Dictionary of Contemporary English). Accessed on 20 May 2021.https: // www. ldoceonline.com/.

MEDAL (Macmillan English Dictionary for Advanced Learners). Accessed on 20 May 2021. https: // www. macmillandictionary.com/.

MWCD (Merriam-Webster's Collegiate Dictionary). Accessed on 20 May 2021. https:// www.merriam-webster.com/.

NOW (News on the Web). Accessed on 20 May 2021.https://www.english-corpora.org/now/. OALD (Oxford Advanced Learner's Dictionary). Accessed on 20 May 2021. https:// www. oxfordlearnersdictionaries.com/.

OED (Oxford English Dictionary). Accessed on 20 May 2021. https://www.oed. com/.

B. Other Literature

Bejan, C. (2017). *English words: structure, origin and meaning*. New York: Addleton Academic Publishers.

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. Cambridge: Cambridge University Press.

Bussmann, H. (1996). *Routledge dictionary of language and linguistics* (trans. and eds.) G. Trauth & K. Kazzazi. London/New York: Routledge.

Crystal D. (2008). *A dictionary of linguistics and phonetics* (6th Edition). New Jersey: Blackwell Publishing.

Edwards, G. (1968). *Names became everyday words, uncumber and pantaloon: Some words with stories.* Holt & Company London: Geoffrey Bles.

Fromkin, V., Rodman, R. & Hyams, N. (2003). *An introduction to language* (7th ed.). Boston/ Massachuset: Wadsworth.

Gao, Y. W. (2012). *Caught in the web of words: Research into English neologisms and dictionaries*. Shanghai: Fudan University Press.

Hendrickson, R. (1972). *The dictionary of eponyms*. New York: Stein & Day.

Lalic-Krstin, G. (2004). Eponyms in English. *Romanian Journal of English Studies*, 1, 69-74.

Landau, S. I. (1984). *Dictionaries: The art and craft of lexicography*. Cambridge: Cambridge University Press.

McArthur, T. (1996). *The oxford companion to the English language*. Oxford: Oxford University Press.

Minkova, D., & Stockwell, R. (2009). *English words: History and structure*. Cambridge: Cambridge University Press.

Mufwene, S. S. (1988). Dictionaries and proper names. *International Journal of Lexicography*, 1(3), 268-283. https://doi.org/10.1093/ijl/1.3.268

Štekauer, P. (1997). On the semiotics of proper names and their conversion. *Arbeiten aus Anglistik und Amerikanistik*, 22 (1), 27-36. ·

Ten Hacken, P. (2006). Word formation in an electronic learners' dictionary: ELDIT. *International Journal of Lexicography*, 19 (3), 243-256. https://doi.org/10.1093/ijl/ecl012

Urdang, L. (1996). The uncommon use of proper names. *International Journal of Lexicography*, 9 (1), 30-34. https://doi.org/10.1093/ijl/9.1.30

CONSTRUCTION OF A COMPARATIVE DICTIONARY OF SINITIC AND SINOXENIC LANGUAGES COGNATES PHONOLOGY

Louis Lecailliez
Graduate School of Informatics, Kyoto University, Japan
louis.lecailliez@outlook.fr

Abstract

About a dozen languages in East-Asia share an important number of cognates because of a common origin (Sinitic family) or extensive borrowings (Sinoxenic languages). This is a useful fact for a speaker who masters one of them and want to learn another. In a bilingual or multilingual dictionary, lexicographic information can be compared but the burden of analysis is placed on the user. This paper describes the construction of a dictionary of comparative phonology of cognates in Sinitic and Sinoxenic languages that targets learners of any of the languages it contains (presently Japanese, Standard Chinese, Taiwanese Southern Min and 6 Hakka dialects). The main dictionary's goal is to make explicit phonological similarities and differences in synchrony between cognates and teach non-obvious phoneme correspondence rules in-between those languages. We expose the theoretical framework and detail the relevant issues and their solutions. In particular, the level of representation (phonetic vs phonemic) and the implication of considering the union set of phonemes of multiple languages are discussed. Practical issues such as dealing with the different scripts and romanizations are also addressed. A comparison algorithm derived from the method of consonant classes from historical comparative linguistic is presented. Finally, we illustrate the planned output with the current prototype of an entry, which make use of the comparison algorithm for displaying data. We conclude on possible future derivate works, enabled by the digital nature of the project, that is fully automated and relies on open-data lexical resources.

Keywords: cognates, learner's dictionary, comparative phonology, multilingual dictionary, language learning

1 Introduction

Learning a language is an activity that can yield numerous benefits on professional and personal levels. In East-Asia, cultural phenomena such as the Chinese literary classics, Japanese animation or Korean popular music are powerful factors that drive people to start learning a language. Migrations and business considerations are other circumstances driving millions to learn an additional language. Moreover, interest in those languages exist in the rest of the world of well.

The task of learning a language is however not a small task and it takes a considerable amount of time and efforts to reach a stage of useful proficiency. Any time and effort spared can be re- invested in advancing to a better proficiency and lower the probability of the learner to give up. In the case of Sinitic and Sinoxenic languages (see the next two sections for a definition), there is an important number of shared cognates, that is "a linguistic form which is historically derived from the same source as another form" (Crystal, 2011). However, sound changes that occurred in each language, as well divergences in their phonology and writing system have obscured their similarity.

A dictionary of cognates would expose the proximity of pronunciation in-between languages and help fostering a multilingual environment both inside one country, and in relation to others, without having to resort to a very distant language such as English. This dictionary would explicit how the pronunciations of cognates relate to each other in different languages, which would help a learner transfer the lexicon he already. This article describes the creation of a comparative dictionary of East-Asian cognates phonology (東亞語言發音對照辭典*, Dōngyǎ yǔyán fāyīn duìzhào cídiǎn*) which aims to support that use case.

More precisely, the dictionary goals are to help learners re-use vocabulary by making explicit sound correspondences between cognates of Sinitic origin, promote multilingualism by including many languages and dialects, and provide a re-usable framework and data for future research supporting the two previously stated goals.

1.1 Sinitic Languages

The Sinitic family of languages (Handel, 2015) is part of the wider Sino-Tibetan family. It regroups a variable number of languages, depending on the linguist describing it. One of these classifications (Kwok, 2018) lists: Mandarin (官話 *guānhuà*[1]), Wu (吳語 *wúyǔ*), Yue (粵語 *yuèyǔ*) also known as Cantonese, Min (閩語 *mǐnyǔ*), Xiang (湘語 *xiāngyǔ*), Hakka (客家話 *kèjiāhuà*), Gan (贛語 *gànyǔ*), Jin (晉語 *jìnyǔ*), Hui (徽語 *huīyǔ*) and Pinghua 平話 (*pínghuà*). All these languages share traits such as being tonal languages and having a common syllable structure (Wee & Li, 2015).

1.2 Sinoxenic Languages

The so-called Sinoxenic (Martin, 1953) languages do not form a single family. Instead, the term designates languages which share the common characteristics of having heavily borrowed vocabulary from Middle Chinese; Late Middle Chinese in the case of Korean (Lee, 1994). The languages in question are Japanese, the larger representant of the Japonic family — its other sub-family being formed by Ryukyu languages —, Korean from Koreanic family that also includes Jeju language, and Vietnamese which is part of the Austra-Asiatic > Mon-Khmer > Viet-Muong family hierarchy (Eberhard, Simons, & Fenning, 2021). It is the important amount and systemic borrowings from Chinese that distinguish Sinoxenic loadwords from sporadic and earlier borrowings (Sybesma et al., 2017). For instance, the word *ume* (梅) in Japanese, coming from Old Chinese **hmay*, is not considered a Sinoxenic borrowing since it was done earlier than the systematic borrowing period and done in isolation.

In Japanese, borrowings that happened during Middle Japanese (Early Middle Japanese: 800-1200, Late Middle Japanese: 1200-1600) from Chinese was so substantial it is qualified a “sinification” of the language by Frellesvig (2010). The number of loanwords was so considerable it brought new phonological phenomena to the language such as palatalization (Labrune, 2016) and bent the existing rules of the language that forbid /r/ at word initial (Labrune, 1993).

The systematic Sinoxenic borrowings include the borrowing of the Chinese writing system and a large corpus of texts, notably the Classics and religious literature (Buddhism). Since the Chinese characters weren’t adapted to write non-Sinitic languages, all the Sinoxenic cultures first used Classical Chinese as the language of written communication, then developed a way to write their vernacular language. Vietnamese used a combination of Chinese characters and characters coined on the model of Sinograms called *chữ nôm* for around a millennia before switching to a script based on the Roman alphabet (Phuong, 1978).

[1] In this paper, words will be glossed in Standard Chinese with *hanyu pinyin* by default, even when the word exist in other languages.

2 Related Work

2.1 Research

2.1.1 Theses on Multilingual Knowledge Transfer

Two recent doctoral theses defended in France, (Labbé, 2018) and (Goudin, 2017) disserted the transfer of knowledge from a known language to another of the same family. Labbé's work dealt with West and South-Western Slavic languages. The section 2-4 is dedicated to underlining the importance of orthographic and phonological equivalence in vocabulary, which stems from historical phonology, where he argues that those can be presented in a "synchronic fashion". This is the approach taken by the dictionary presented here: while historical phonology phenomena are the source of the existing phoneme correspondences in synchrony, making a learner study a reconstructed language and sound change laws to understand current phonological correspondences is adding a huge burden to his learning. The goal of the dictionary is to lower the amount of work for the student, not to double it, so historical reconstructions and the applicable sound changes are explicitly out of the scope of this project.

Goudin's thesis is more directly appliable to the present work since it is a reflection on the use of Sinograms (Chinese characters) as a tool for inter-comprehension between Standard Chinese, Korean and Japanese. Sadly, it is hard to know more about since the thesis isn't available online. The main difference however, is that the Chinese character is the basic unit of analysis, with radicals and pronunciations being the sub-unit of analysis. In the present work, the basic unit being listed is the lemma, with the sub-unit being the syllable.

2.1.2 Contrastive Database of Japanese and Taiwanese Pronunciations

Nakazawa, Iwaki & Koresawa (2013) constructed a comparative table of pronunciation of Chinese characters in Taiwanese Southern Min and Japanese based on the 日台大辞典 (*nittai daijiten*, Japanese-Taiwanese Grand Dictionary) dictionary. Another database was created by Sakai & Nakazawa (2017), which is based on the content of the 台日新辞書 (*tainichi shinjisho,* Taiwanese-Japanese New Dictionary). Both projects have for stated goal to help Taiwanese learners of Japanese and spread the awareness in Japan of the fact that pronunciation of *kango* (漢語, Sino-Japanese words) are more similar to Japanese in Taiwanese than in Standard Chinese. Both databases are available for download as Excel files.

The present dictionary shares the goals expressed in those two papers. The biggest difference lies in the basic unit of comparison, which is the lemma in the cognate dictionary and the Chinese character in the Japanese-Taiwanese comparison table and database. In addition, while Nakazawa et al. (2013) mentions Hakka, Cantonese Vietnamese and Korean as possible future extension of their database, Hakka is integrated from the start in the dictionary presented here and resources have been collected for the three other languages. The technical mean of distribution differs too: Excel file in one hand, a website on the other hand.

2.1.3 Research on Semantic Comparison between Japanese and Chinese

By their prominence in the Japanese language, *kango* have attracted attention of linguists and lexicographers and some works classified the proximity of those words in-between Japanese and Chinese on the semantic level.

Matsushita et al. (2017) developed a database of Japanese-Chinese *kango* comparison. The resulting database is freely accessible on the web. The database lists semantic correspondence patterns such as same, overlapping, or different meaning of the cognate pairs. Xiong & Tamaoka (2014) analyzed the semantic similarity of words made of two characters and found that ~60% of the pairs share the same exact meaning, and an additional ~29% Japanese *kango* have all the Chinese meanings, in addition to Japanese specific ones. On a larger set of 20,000 lexemes, Matsuhita et al. (2017) found a very similar percentage for the noun category: 62.3% of the *kango* and their Chinese counterpart have an identical meaning.

From those research results, it is clear that the difference of meaning in cognates will not be too problematic in the general case and that an important number of cognates are easily transferred on the semantic level. Difference in semantic is thus addressed well in research literature and in the dictionary landscape while phonology isn't. In particular, two comparative dictionaries of Japanese and Chinese have been published, one using words as entries (Wang, Xu & Kodama, 2007) and the other listing Chinese characters (Tang, 1993).

2.2 Dictionaries

2.2.1 Trilateral Cooperation Secretariat Dictionaries

The trilateral Cooperation Secretariat published a set of three dictionaries (one in Japanese, one in Chinese and one in Korean) which list 658 Chinese words. For each entry, the writing in Chinese character is given (Simplified Chinese is used), their pronunciation in *romaji* (Latin letters), *hanyu pinyin* and hangul. At least one meaning is given for an entry, which is accompanied by multiple examples given in the three languages. Each example has the same meaning. However, nothing is done in those dictionaries to explicit the correspondence or divergence of pronunciations of words in-between the three languages.

2.2.2 Proto-Indo-European Lexicon Dictionary

The Proto-Indo-European Lexicon (Pyysalo et al., 2019) has the particularity of not containing directly dictionary entries for the languages it aims to support. In fact, that would be very difficult to do given that 150 to 200 languages are projected to be included. Instead, each language encodes sound change laws with a computer technology (finite-state automaton). Entries in attested languages are generated from the PIE roots by applying successively every sound change rules; when the results divert from the attested form, it is highlighted in red the presentation. The focus is "initially" placed on etymology and more information are provided by linking to existing dictionaries present on the web.

This project, in its technical execution is very similar to the one presented here: both are starting from a small set of third-party data and are encoding linguistic facts as code to make transformations on a set of starting lexicographic data. The data displayed are for the most part computed. Comprehensive lexicographic information (such as meaning) for each language is delegated to existing dictionaries by linking to them. Presentation of data is highly customizable in the interface, albeit not all features are implemented yet.

2.2.3 German-English Etymology Dictionary

Qu (2007) describes an etymology dictionary for Chinese learners of German that have a good command of English already. The stated goal is to allow users to recognize cognates in-between German and English despite the fact "phonological and semantic evolution has concealed much of their formal similarity" and thus allow them to leverage their existing knowledge of English. In contrast to Sinitic and Sinoxenic languages, the sound changes have been more radical in Germanic, leading to cognates that significantly diverge in pronunciation and orthography. Both the Old High German (OHG) and Old English (OE) words are given for a cognate pair, making their relationship more obvious. For example (Qu, 2007): "day (<OE. dæg) – Tag (<OHG. Tag)". In addition to phonology, the dictionary gives semantic information: signposts are used to warn users about important divergence in meaning. The common point of Qu's work and the

present dictionary is that both recognize the importance of phonology of cognates for transferring existing knowledge of a subset of vocabulary to another language.

3 Methodology

Similarly, to the PIE Lexicon project, the dictionary presented here is not a dictionary produced in the traditional fashion: there are no lexicographers or users writing entries. Instead, the content of existing dictionaries is reused, transformed and aggregated to provide new functionality absent from the original dictionaries. The value of the present work lies in aggregating information from disparate sources and the highlighting the similarity and difference in cognates pronunciation in-between different languages.

3.1 Project Overview

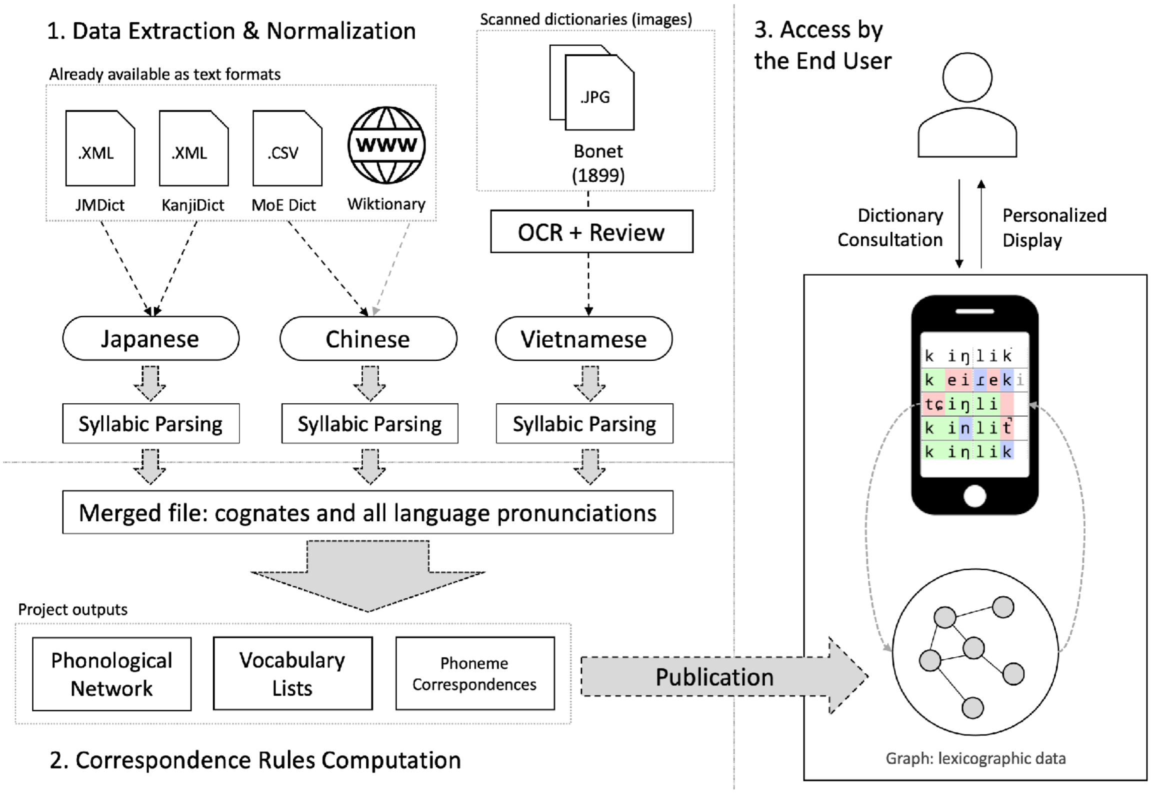


Figure 1: Project technical architecture

The project is structured as collection of data files (see Figure 1, *1. Data Extraction & Normalization*) used as input for a subsequent processing chain (*2. Correspondence Rules Computation*). The data are mainly composed of dictionaries published under open-data licenses such as JMdict (Breen, 2004), but additional resources that have a pedagogical value are used as well. Once data for a language have been collected, they are normalized to fit a common syllable format (see Section 4.3). All normalized word pronunciations are then regrouped under their cognate written in traditional Chinese characters ("Merged file" on Figure 1). The extraction and normalization phase can be arbitrary complex and is done with a collection of programs and scripts written for this purpose.

The merged file is the starting point of the different planned outputs of the project. The main output is a lexical network which is exposed to the public through a website using an existing software platform developed for another dictionary research project (Lecailliez et al., 2020). The website, which will support mobile consultation, is still under development.

An important principle of the project is the requirement that all its output can be recreated from the original data and the transformation chain. This ensure: (1) new and updated data can be retrieved from the original dictionary projects when those get updates, (2) errors introduced by the processing chain can be corrected by fixing the code involved and rebuilding the whole project and (3) the output of the project is parameterizable, which allows for different outputs based on different linguistic modeling.

3.2 Data Sources

Table 1 lists the dictionaries (column 2) used as data sources by the project for each language. The last column indicates how many entries are extracted from the source. When multiple dictionaries were collected for a language, an asterisk (*) marks the dictionary from which entries are extracted. In the case of Sinoxenic languages, the percentage indicates the proportion of extracted entries (for Sinitic languages almost every entry is extracted).

Table 1: Dictionaries used as data sources

Language	Dictionaries	Extracted Entries
Mandarin	重編國語辭典修訂本	160,658
Cantonese	CC-CANTO*, Cantonese Wordnet	105,862
Japanese	JMDict*, KanjiDict	75,351 (~66.7%)
Taiwanese	臺灣閩南語常用詞辭典, 台日大辭典*	56,466
Korean	Kengdic	38,255 (~28.6%)
Hakka	臺灣客家語常用詞辭典	14,484
Vietnamese	Dictionnaire annamite-français, Wiktionary*	5212
Central Okinawan	沖縄語辞典	2,236 (~15,4%)

These dictionaries have been created using different methodologies. Most have been complied be a team of lexicographers or linguists (in particular the ones from Taiwan) while some are crowd-sourced (JMDict (Breen, 2004), KanjiDict, Kengdic). Both CC-Canto and the Cantonese Wordnet (Sio & Morgado da Costa, 2019) employed native speakers to check the pronunciation of words. The sources are thus generally highly trustable, especially since only the pronunciations are extracted, which limit the surface of possible lexicographic issues and the problem of combining dictionaries compiled using different methodologies.

Since the Vietnamese-French dictionary (*Dictionnaire annamite-français*, 大南國音字彙合解 大法國音 *Đại Nam quốc âm tự vị hợp giải Đại Pháp quốc âm* (Bonnet, 1899)) is only available as scanned images it doesn't fit the existing processing chain and entries are extracted yet. Given the complexity of the task (Lecailliez, 2015) this part will likely need to be done manually. Japanese requires the use of a Chinese character dictionary for parsing its words readings unambiguously hence the inclusion of a kanji dictionary (KanjiDict). The Hakka dictionary contains dialects of 6 locations (四縣 *Sìxiàn*, 海陸 *Hăilù*, 大埔 *Dàbù*, 饒平 *Ráopíng*, 詔安 *Zhàoān*, 南四縣 *Nánsìxiàn*), each of them having them than 13,000, entries save for the Zhaoan dialect which contains only 10,508 words.

Licenses of those dictionaries varies from freely reusable even commercially, to copyright free, passing by allowing reuse without modifications. Most licenses involved are a Creative Commons one. Some are incompatibles with each other or does not allow modifications to be distributed. In particular the 重編國語辭典修訂本 (*zhòng biān guóyǔ cídiǎn xiūdìng běn*, Revised Chinese Dictionary) is available to download and allows reproduction but does not allow redistribution of derivative works. Rights of use will need to be negotiated with copyright holders to make use of the content of those dictionaries.

Three kind of additional information are relevant to the project: semantic comparison, frequency and pedagogical levels. The only file collected so far about semantic comparison is the database created by Matsushita et al. (2017). Wiktionary provides frequency lists for an important number of languages. Pedagogical levels refers to level of standard tests like the JLPT (日本語能力試験, *nihongo nōryoku shiken*), HSK (漢語水平考試, *hànyǔ shuǐpíng kǎoshì*) or TOCFL (華語文能力測驗, *huáyǔwén nénglì* *cèyàn*) when they exist for a language. In some case official vocabulary level lists are available, for other lists have been compiled by netizens. Since they all use different rating, a standardization based on CEFR levels is done. Those data will be used for outputs and features that are outside the scope of this paper.

4 Linguistic Modeling

4.1 Script Normalization

The sources dictionaries make use of 6 different scripts: Chinese characters, katakana, hiragana, hangul, *zhuyin fuhao* (also called bopomofo) and Latin script. Roman alphabet is used for very different romanization schemes: *tâi-lô and peh-ōe-jī* for Taiwanese, *jyutping* for Cantonese and the Vietnamese alphabet. All differ in the value they assign to letters. While it is reasonable to expect the reader to be able read one or two writing systems, it is unrealistic to expect the average user to know the intricacies of a dozen scripts and romanizations. In particular since entries juxtapose pronunciations of a word in multiple languages, confusion in letters' value could arise easily. To solve this issue, the readings of cognates are transformed from their original script to the International Phonetic Alphabet (IPA). Choosing the IPA doesn't solve all problems however: the transcription used could be either phonological or phonetic.

4.2 Phonemic vs Phonetic Transcription

This dictionary lists how cognates are pronounced in the languages it includes. The IPA alphabet is used for that task, but it raises the question of using a phonemic or phonetic transcription. Generally speaking, a phonetic transcription contains more information than a phonemic one. It makes them harder to read for a non-trained user and requires precise information that are present only in specialized dictionaries. The present dictionary thus leans towards phonological transcriptions.

The use of phonological transcription is however problematic in a multilingual context because the phonological system of a language abstract differences that can be meaningful in another language; this occurs particularly with contextual allophones. For instance, Japanese /s/ in front of /i/ is realized as [ɕ] (Labrune, 2006, p. 81). As the same phenomenon applies in Korean (Shin, Kiaer & Sha, 2017, p.70) this is not an issue when comparing words in those two languages. It is however a problem when Japanese is compared to Chinese where both /s/ and /ɕ/ have phonemic status. The same phenomenon applies even if two allophones doesn't exist per se in a language known by a learner but match close ones. For instance, /h/ in Japanese has [h] and [ɸ] as contextual allophones. The phoneme /h/ does not exist in French while [ɸ] would be easily interpreted as /f/.

The way the cognate dictionary handles this problem is to use a phonological transcription that distinguish contextual allophones when relevant (one a case-by-case basis).

4.3 Syllable Structure

Sinitic languages share a common syllable structure made of at most four segments (Wee & Li, 2015). This pattern is commonly referred to as CGVX where C is a consonant, G a glide, V the main vowel and X the coda which can be either a consonant or a vowel. Any segment except the main vowel one is optional. An alternative syllable pattern is that of a single syllabic consonant. The syllable can be described as a tree, for which competing theories exist. For this project the hierarchical model does not yield benefit and a flat model is used instead. In addition, each syllable possesses a tone. An exception to the model exists in Standard Chinese because of the *erhua* (兒化) phenomenon; it is currently not handled by the dictionary and the few entries affected are discarded.

Vietnamese and Korean syllables fit the pattern as well. Japanese exhibits an epenthetic vowels /u/ or /i/ after -/k/ and -/t/ coda. This vowel is discarded for phonological comparison to other languages but is displayed to the user. In the dictionary, diphthongs are split in two parts to ease comparisons between languages, the first part is allocated to the main vowel slot while the remaining part fills the coda slot.

4.4 Comparison Algorithm

4.4.1 Slot Comparison Values

An important part of the project is the similarity algorithm it defines. Phonetic similarity is used in various works pertaining to Chinese Natural Language Processing (NLP); we can cite (Chang et al., 2010) and (Lee et al., 2019) as examples. Metrics created for those works are tailored to the task at hand, and offer limited reusability for a different purpose. Since no existing algorithm fitted our goal, a new one was devised. A measure of similarity between two syllables will allow searching similarly sounding syllable across languages, and provide a numeric value to sort vocabulary, for instance when creating vocabulary lists.

The metric must work across languages, be close to human judgment that is if a human would judge two syllables very similar the score should be very high and it must be computable from the data extracted from dictionaries (i.e. we cannot afford to measure the actual perception in- between all the speakers of the languages involved).

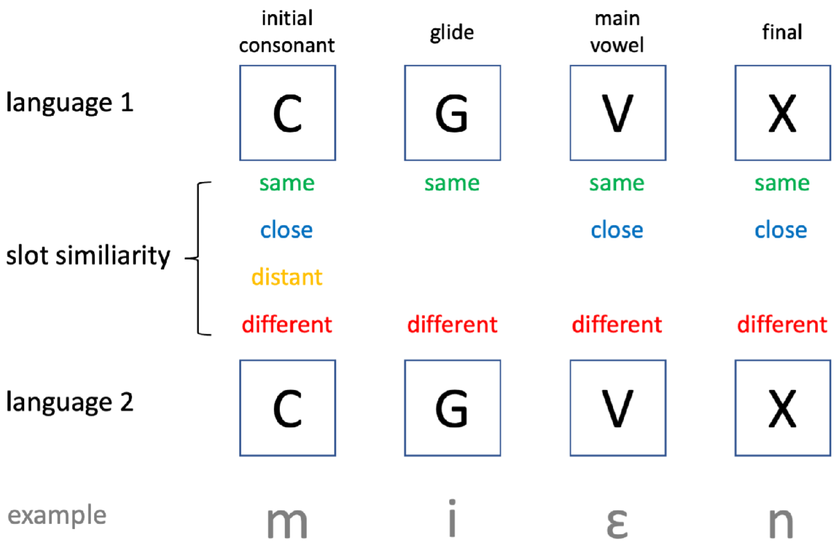


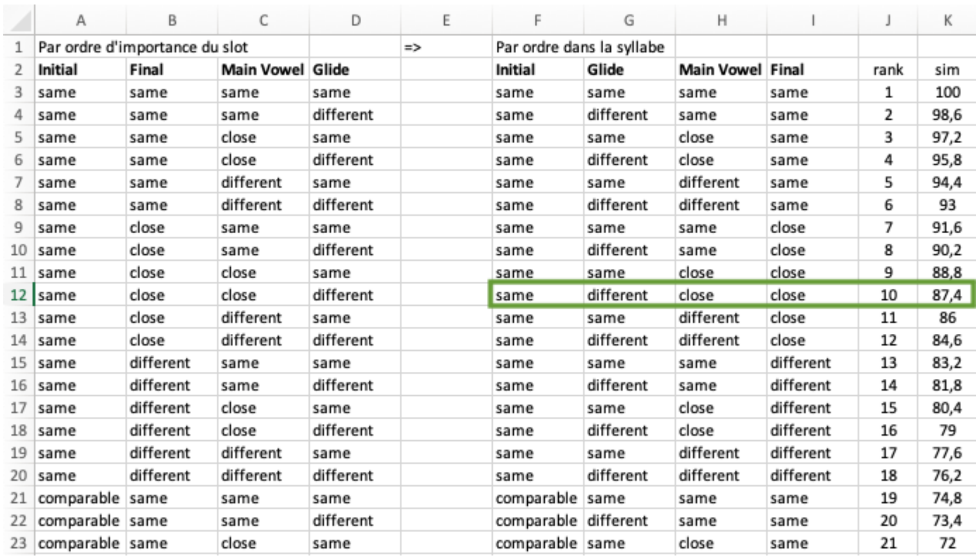
Figure 2: Syllable slots and possible comparison values

The algorithm works by comparing each pair of slots. If the phoneme is identical, the output for the slot is the value "same". If the phonemes are somewhat close, which is determined on the basis of the user native language and feature geometry (see below), the output for the slot is "close". Otherwise, the output is "different". Since the initial consonant is the part of the Sinitic syllable where is the more variety an additional "distant" output value exists. At the syllable level, the number of resulting output combinations is 72 (4*2*3*3).

Intuitively each slot doesn't participate in the same weight in the similarity between two syllables: for instance, the glide can be absent in one of them without making the syllables too different. More importantly, consonant information is more impactful than vowel one as confirmed to its relative stability over time, and across places and language borrowings (which make the present work feasible in the first place) while vowel information is often highly variable even within dialects of the same language. Those, the algorithm prioritizes consonant information and use the following order of slots: initial, final, main vowel, glide.

4.4.2 Ranking and Similarity

The 72 possible combinations are constructed from the most similar (same, same, same, same) to the most dissimilar (different, different, different, different) and are each effected a rank ranging from 1 to 72. Since metrics usually range from 0 to 1 or 0 to 100, the rank is converted to a measure ranging from 0 to 100 by using the formula *floor(100-(rank-1)*1.4)*.



	A	B	C	D	E	F	G	H	I	J	K
1	Par ordre d'importance du slot				=>	Par ordre dans la syllabe					
2	**Initial**	**Final**	**Main Vowel**	**Glide**		**Initial**	**Glide**	**Main Vowel**	**Final**	rank	sim
3	same	same	same	same		same	same	same	same	1	100
4	same	same	same	different		same	different	same	same	2	98,6
5	same	same	close	same		same	same	close	same	3	97,2
6	same	same	close	different		same	different	close	same	4	95,8
7	same	same	different	same		same	same	different	same	5	94,4
8	same	same	different	different		same	different	different	same	6	93
9	same	close	same	same		same	same	same	close	7	91,6
10	same	close	same	different		same	different	same	close	8	90,2
11	same	close	close	same		same	same	close	close	9	88,8
12	same	close	close	different		same	different	close	close	10	87,4
13	same	close	different	same		same	same	different	close	11	86
14	same	close	different	different		same	different	different	close	12	84,6
15	same	different	same	same		same	same	same	different	13	83,2
16	same	different	same	different		same	different	same	different	14	81,8
17	same	different	close	same		same	same	close	different	15	80,4
18	same	different	close	different		same	different	close	different	16	79
19	same	different	different	same		same	same	different	different	17	77,6
20	same	different	different	different		same	different	different	different	18	76,2
21	comparable	same	same	same		comparable	same	same	same	19	74,8
22	comparable	same	same	different		comparable	different	same	same	20	73,4
23	comparable	same	close	same		comparable	same	close	same	21	72

Figure 3: The first of the 72 possible comparison combinations and their ranks

Figure 3 illustrates how the ranks are computed. On the left the natural progression of ranks is visible (slots are sorted by importance), on the right the slots are re-ordered corresponding to their actual position in the syllable. The two leftmost columns display the rank and the associated similarity. For words, the similarity score is computed using the geometric mean of each syllable similarity. In comparison to the more common arithmetic mean, the geometrical mean is more sensible of important gap in value (e.g. the geometric mean of 1 and 100 is 10).

For example, Japanese 愛 (ai, *love*) and Chinese 愛 (ài, *love*) share the same initial and glide (both empty) as well as the same main vowel and final one. The algorithm gives them a rank of 1, equating a similarity of 100. The Chinese 麵 (*mi*àn, noodle) and Japanese 麺 (*men*, noodle) have the same initial, a different glide, a close main vowel and a close final, leading to a rank of 10 which give a similarity of 87 (see line highlighted in green on Figure 3).

4.4.3 Consonant Comparison with Language Profiles

The comparison of consonants is inspired from the method of consonants classes initiated by Dolgopolsky (1986) and used in comparative-historical linguistics. Examples of such classes can be found in (Kassian et al., 2015). The class of labials (P-class) for instance contains the consonants: p b β ɓ f v… Those classes however are too broad for use in this project.

Another difference is that the data in comparative linguistics are absolute. However, the perception of a phoneme from a foreign language depends on one's native language.

Table 2: Presence and absence of phonemes /k/, /kʰ/, /g/ in Chinese, Japanese, Taiwanese and French

Phoneme	**/k/**	**/kʰ/**	**/g/**
Chinese	/k/	/kʰ/	—
Taiwanese	/k/	/kʰ/	/g/
Japanese	/k/	—	/g/
French	/k/	—	/g/

One of the common error of speakers (Teramura, 1990) having Chinese as a native language who are learning Japanese as a second language is with the voiced/devoiced characteristic of bilabial plosives (/b/, /p/), alveolar plosives (/d/, /t/) and velar plosives (/g/, /k/) which stems from the voiced series not existing in Chinese. Thus, upon hearing a Japanese word containing a voiced consonant that consonant may be mistaken for its unvoiced counterpart. On the contrary, a native speaker of Taiwanese or French for which the distinction exist will be able to recognize that phoneme correctly. This phenomenon calls for using finer consonant classes, and a different mapping from phonemes to classes that depends on the language of the reader, and on the ability to discriminate phonemes in the second language he is learning.

The output “close” and “distant” is realized in the comparison algorithm by affecting to each phoneme of a language a given class and seeing if the classes match. The association of phonemes to classes is done for every language of the expected readers of the dictionary (this work can be crowd-sourced). For instance, both of the Japanese phonemes /k/ and /g/ are mapped to the class K in the “close” profile language for native speakers of Chinese beginner in Japanese, while /k/ is mapped to K and /g/ to G in the “close” profile of Taiwanese, Japanese, French speaker and advanced learner of Japanese. Moreover, both /k/ and /g/ are associated to class K in “distant” profiles of Taiwanese, Japanese and French. Thus, when comparing 乾 (Chinese *gān*, Japanese *kan*, dry) in Chinese and Japanese the initial will be rated as “close” (since both as K-class) from the point of view of a native Chinese-speaker beginner in Japanese, while being rated only “distant” for a Taiwanese, Japanese, French speaker or advanced learner.

4.5 Correspondences Rules

Besides a visually compelling table of phoneme-to-phoneme comparison, the dictionary aims to include regular correspondences rules between phonemes in language pairs. Despite parallel language evolution, phonemic correspondences still exist in-between the languages included in the dictionary. Some of those correspondences are obvious such as /f/ in Chinese and /h/ in Japanese (方法 *hōhō / fāngf*ǎ, method) while others are less evident; for example Chinese nasal coda -/ŋ/ is usually found as a long vowel in Japanese (e.g. 方 *fang / hō*, direction).

The data and processing tools in the project have for goal to found those correspondences in- between any language pairs, and to compute statistics about their frequency, regularity and their pedagogical potential.

To give an illustration of correspondence rules and their application, let’s observe the pronunciation of three morphemes in Japanese and Vietnamese. For each morpheme, the table 3 give first the pronunciation of the morpheme (in *romaji* for Japanese and *quốc ngữ* for Vietnamese) and then lists a simplified phonemic representation where the glide and vowel information are discarded (symbolized by _). An empty coda is noted ø.

Table 3: Vietnamese and Japanese pronunciation of three morphemes

Morpheme	**Vietnamese**		**Japanese**	
言	ngôn	ŋ _ _ n	gen	g _ _ N
語	ngữ	ŋ _ _ ø	go	g _ _ ø
我	ngã	ŋ _ _ ø	ga	g _ _ ø

From the data listed in Table 3, it is possible to infer the rules listed in Table 4.

With knowledge of these rules, a speaker of Vietnamese will be able to infer that a morpheme pronounced *nguyen* will have the shape g _ _ N in Japanese. Indeed, the rule is true for 元 and 原, both pronounced *nguyên* in Vietnamese and *gen* in Japanese. For those morphemes, the only additional information that a learner has to memorize is the main vowel and glide values. The burden of learning is reduced in comparison to a learner without any prior knowledge.

Table 4: correspondence rules inferable from Table 3

Rule	Representation
Vietnamese voiced velar nasal /ŋ/ at initial is found as voiced velar plosive /g/ in Japanese	ŋ _ _ _ → g _ _ _
Vietnamese -/n/ coda is found as -/N/ coda in Japanese	_ _ _ n → _ _ _ N
Vietnamese empty coda corresponds to an empty coda in Japanese	_ _ _ ø → _ _ _ ø

This is only a small example and rules present in the dictionary will be extracted using all the available data. The important number of cognates pairs collected for most pair of languages (see Table 5) allows to compute how frequent and regular the correspondences are over the lexicon. Correspondence rules will be listed under every entry they are appliable to.

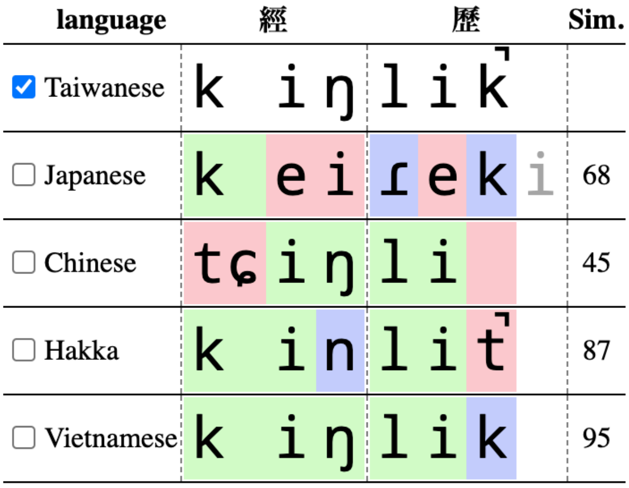
5 Results

5.1 Visualization

The algorithm detailed in Section 4.4 can be used to produce a colored visualization of the difference of pronunciation of a cognate between multiple languages. A graph visualization involving all the language pairs would be hard to read, so the adopted solution is to display comparison data as a dynamic table: one language serves as the basis of comparison and that language can be changed by the user.

The Figure 4 shows the table generated for the cognate (經歷, *jīnglì*), with Taiwanese being used as the basis of comparison. The selected language is put as the first row of the table, and its corresponding checkbox is ticked. In addition, since it's the basis of the comparison, none of its phonemes are colored. The remaining rows of the table contains the other language pronunciations, each with phoneme slots colored based on the output of the comparison algorithm with the top row (see Figure 2 for the meaning of colors).

The content of individual cell is padded with white space so the initials, medials, central vowel and finals are always aligned regardless of their length. A monospace font is used to ensure the alignment is possible. When a slot is missing in every language (the medial of each syllable in the example), the slot is removed from display not to clutter the table with empty columns. The epenthetic vowel present in Japanese is displayed as an addition slot, greyed to indicate the special nature for the vowel in respect to the common syllable structure.

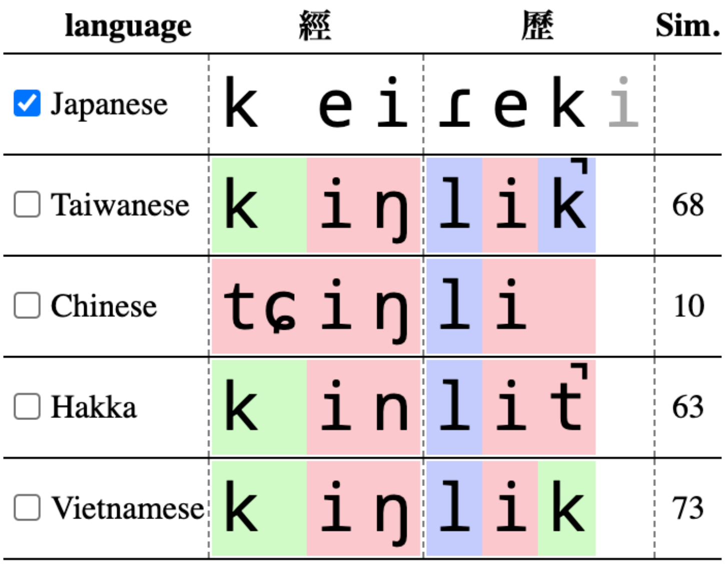


language	經	歷	Sim.
☑ Taiwanese	k iŋ	lik̚	
☐ Japanese	k ei	ɾeki	68
☐ Chinese	tɕiŋ	li	45
☐ Hakka	k in	lit̚	87
☐ Vietnamese	k iŋ	lik	95

Figure 4: Entry "經歷" with Taiwanese as comparison basis

This visualization makes very explicit which phonemes are identical in other languages. Moreover, it also gives a quick impression of how the cognate differs in comparison to the other languages: in this case the Taiwanese pronunciation is quite close to most of other languages. In addition, the similarity ("Sim.") column gives the numeric computation of the closeness of pronunciation, which can be used to infirm or confirm the impression given by the coloring scheme.

The base language for comparison is changed by ticking the checkbox corresponding to another language. The Figure 5 displays the same data (經歷 cognate) but uses Japanese as the basis of the comparison. It is immediately clear that the Japanese pronunciation differ greatly from the all languages by the number of red cells present in the table. Besides the initials of the two syllables with is identical or close to most other languages (except Standard Chinese), every other slot, save for the final in Vietnamese, differs.



language	經	歷	Sim.
☑ Japanese	k e i	ɾ e k i	
☐ Taiwanese	k i ŋ	l i k̚	68
☐ Chinese	tɕ i ŋ	l i	10
☐ Hakka	k i n	l i t̚	63
☐ Vietnamese	k i ŋ	l i k	73

Figure 5: Entry "經歷" with Japanese as comparison basis

5.2 Shared Vocabulary Between Languages

The dictionary wouldn't be of effective utility if there wasn't a significant number of cognates shared by the languages involved. Since data have been extracted for 7 languages, it is possible to compute the vocabulary common to the possible language pairs (that is, the intersection of their vocabulary).

Table 5 lists the vocabulary in common for the top-6 languages in terms of vocabulary size included in the project. Languages are listed in the first or second column based on the number of entries extracted for that language, the one having the bigger number being put on the first column.

Table 5: Vocabulary common to language pairs

Language 1	Language 2	Shared Cognates
Mandarin	Cantonese	54,024
Mandarin	Taiwanese	19,496
Mandarin	Japanese	18,120
Cantonese	Taiwanese	15,025
Cantonese	Japanese	14,843
Japanese	Korean	11,552
Mandarin	Korean	9,856
Mandarin	Hakka	9,318
Cantonese	Korean	8,630

Japanese	Taiwanese	8,369
Cantonese	Hakka	8,300
Taiwanese	Hakka	6,596
Taiwanese	Korean	4,808
Japanese	Hakka	3,179
Korean	Hakka	1,987

It is notable that 6 combinations of languages share more than 10,000 words and the majority of the pairs share more than 8000 words. While a lot of this vocabulary may be specialized or of very low frequency, this tends to prove that a speaker or learner will be able to reuse a lot of vocabularies by using the dictionary presented here.

It is also possible to compute the vocabularies that are shared by more than two languages at once. In Table 6, the vocabularies present in sets of 4, 5 and 6 languages are computed. It is remarkable that a relatively high number of words (about 5600) are shared by four languages, including a Sinoxenic one.

Table 6: Vocabulary common to 4-6 languages

Languages	**Shared Cognates**
Mandarin, Cantonese, Japanese, Taiwanese	5,599
Mandarin, Cantonese, Japanese, Taiwanese, Korean	2,574
Mandarin, Cantonese, Japanese, Taiwanese, Korean, Hakka	1,001

In addition, there is a set of ~1000 words that are cognates in 6 languages. Example of such words are: 世紀 (*shìjì*, century), 字典 (*zìdiǎn*, dictionary), 完全 (*wánquán*, complete), 將來 (*jiānglái*, future), 人口 (*rénkǒu*, population), 平和 (*pínghé*, peace), 病院 (*bìngyuàn*, hospital), 論文 (*lùnwén*, article), 中央 (*zhōngyāng*, central), which are useful vocabulary for daily life. Other terms bear special cultural interest: 君子 (*jūnzǐ*, a gentleman in Confucianism), 仙人 (*xiānrén*, an immortal in Taoism); those two words are also present in Vietnamese and Central Okinawan, making them existing in at least 8 languages. In some cases, cognates must be accompanied by an explanation if used in a pedagogical context: 三國 (*sānguó*) refers to different periods in different countries (one in China, one in Korea) and other have a particular meaning in one of the language: 風俗 (*fēngsú*) generally means “customs, traditions” but have the additional meaning of “prostitution” in Japanese.

Those words are “high multilingual” and it is arguable that they are of special interest for a learner of multiples languages. Most vocabulary lists and learning materials are created using frequency and/or educators’ intuition. The multilingual aspect of lexicon can be an objective metric to use as an additional decision criterion for inclusion into a list of vocabulary.

6 Future Work and Conclusion

6.1 Future Work

The main limitation of the present work regards the tonal information of syllables. This information is always extracted from the script and included in the common syllable format used as output but it is not visible in the comparison tables. Tonal information is currently not normalized and uses a conventional number for each language. In the future, the 5 IPA tone levels will be used, which will allow for automatic comparison of tonal information.

The most obvious future development of the dictionary concerns the languages and dialects it includes. In one hand, more Chinese languages such as Wu and Xiang can be added if data is found. Some of the exploitable dictionaries use Chinese characters as entries. At the moment, the dictionary only contains lemma as entries, so this editing choice might be reconsidered. Second, dialectal variation could be integrated into the dictionary. The source dictionary for Hakka already contains such variations. Both

Vietnamese and Korean feature marked difference in-between their Northern and Southern dialectal groups and are interesting target for inclusion. From a technical point of view, adding additional dialects is no different than adding additional languages to the project, and the process is straightforward since the cognate dictionary have been designed for multilingualism from the start.

On a more distant fashion, the dictionary output could be used to produce pedagogical lists of vocabulary. In comparison to existing lists, the lists generated this way could take two variables in account, that are currently ignored: in one hand the proximity of lexical items pronunciation with the equivalent in the learner native (or known) language. In the other hand, the multilingual aspect of a lemma, that is the number of languages in which it exists. More generally, the dictionary is to support comparative work involving the languages it includes. Since the number of possible pairs included is high (15 pairs when considering 6 languages) some of those work may be the first of their kind.

6.2 Conclusion

This paper presented the on-going effort to create a dictionary of Sinitic and Sinoxenic cognates. Dictionaries with satisfying number of entries have been collected for five languages (Standard Chinese, Cantonese, Japanese, Southern Min, Korean), and smaller scale data exist for three others (Hakka, Vietnamese, Central Okinawan). We presented an overview of the processing toolchain use to extract and compute the content of the cognate dictionary. A major contribution of this paper lies in the algorithm created to compute the similarity of syllables across languages. The output of the algorithm is also used to display clearly the difference in pronunciation between words. The algorithm is adaptable to different native language and proficiency of users, which make is usable for other task such as generating list of word a beginner would likely confound.

Finally, we presented quantitative data on the number of shared cognates between 16 language pairs and found that a significant number of cognates are shared across 6 languages, which confirm the potential usefulness of the dictionary. Further research on language transfer as well as generation of vocabulary lists could be made by leveraging the content of the dictionary.

Acknowledgement

I would like to thanks François Lecailliez for funding the conference registration fees.

References

Bonet, J. (1899). Dictionnaire annamite-français: A-M. Leroux.

Breen, J. (2004). JMDict: a Japanese-multilingual dictionary. In *Proceedings of the workshop on multilingual linguistic resources* (pp. 65-72).

Chang, C. H., Lin, S. Y., Li, S. Y., Tsai, M. F., Liao, H. M., Sun, C. W., & Huang, N. E. (2010). Annotating Phonetic Component of Chinese Characters Using Constrained Optimization and Pronunciation Distribution. *International Journal of Computational Linguistics and Chinese Language Processing*, *15*(2), 145-160.

Crystal, D. (2011). *A dictionary of linguistics and phonetics* (Sixth Edition). Blackwell Publishing.

Dolgopolsky, A. B. (1986). A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. *Typology, relationship and time: a collection of papers on language change and relationship by soviet linguists*, 27-50.

Eberhard, D. M., Simons, G. F., & Fennig C. D. (Eds.). (2021). *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.

com.

Frellesvig, B. (2010). *A history of the Japanese language*. Cambridge University Press.

Goudin, Y. (2017). *L'intercompré*hen*sion en langues sinogrammiques: théories, représentations, enjeux, et modalités d'une didactique de la variation*. (Doctoral dissertation, Sorbonne Paris Cité).

Handel, Z. (2015). The classification of Chinese. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 34-44). Oxford University Press.

Kassian, A., Zhivlov, M., & Starostin, G. (2015). Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies, 43*(3-4), 301-347.

Kwok, B. C. (2018). *Southern Min: Comparative Phonology and Subgrouping*. Routledge.

Labbé, G. (2018). *Fondements linguistiques et didactiques de l'intercompré*hen*sion slave : le cas des langues slaves de l'ouest et du sud-ouest*. (Doctoral dissertation, Sorbonne Paris Cité).

Labrune, L. (2006). *La phonologie du japonais* [Version électronique]. Peeters Publishers.

Labrune, L. (1993). À propos d'un trait typologique du japonais: l'absence de r à l'initiale des mots indépendants de Yamato kotoba. *Ebisu*-Étude*s Japonaises*, *2*(1), 7-21.

Lecailliez, L. (2015). *Approches pour une numérisation de qualit*é *d'un dictionnaire vietnamien-français comprenant des caractères Nôm*. (Master's thesis, Paris Diderot, Paris, France).

Lecailliez, L., Flanagan, B., Chen, M.-R. A., & Ogata, H. (2020). Smart dictionary for e-book reading analytics. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (pp. 89-93). Lee, H. (1994). The Origin of Sino-Korean. *Korean Linguistics, 8*(1), 207-222.

Lee, H. (1994). The Origin of Sino-Korean. *Korean Linguistics, 8*(1), 207-222.

Lee, L.-H., Wu, W.-S., Li, J.-H., Lin, Y.-C., & Tseng, Y.-H. (2019). Building a confused character set for Chinese spell checking. In Proceddings of the 27th International Conference on Computers in Education, (pp. 703-705). Asia-Pacific Society for Computers in Education.

Martin, E. S. (1953). The Phonemes of Ancient Chines. *Journal of the American Oriental Society 73*(2), Supplement 16, 1-46.

Nakazawa, N., Iwaki, H., & Koresawa, N. (2013). Possibility of the Japanese-Taiwanese Fundamental Characters' Contrastive Phonetic Table with the Japanese Language Education. 2nd International Conference on Vietnamese and Taiwanese Studies & 6th International Conference on Taiwanese Romanization. National Cheng Kung University, Taiwan.

Phong, N. P. (1978). À propos du Nôm, écriture démotique vietnamienne. *Cahiers de Linguistique Asie Orientale, 4*(1), 43-55.

Pyysalo, J., Kotiranta, F., Sahala, A., & Hulden, M. (2019). Proto-Indo-European lexicon and the next generation of smart etymological dictionaries: The technical issues of the preparation. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 592-602). Lexical Computing CZ sro. Sin, C. Y.,

Shin, J., Kiaer, J., & Cha, J. (2012). *The sounds of Korean*. Cambridge University Press.

Qu, C. (2017). The Necessity of Compiling a Learners' German-English-Chinese Etymological Dictionary. In Proceedings of *ASIALEX 2017: The 11th International Conference of Asian Association of Lexicography* (pp. 483-499). Guangzhou, China.

Sakai, T. & Nakazawa, N. (2017). Database of Holo-Taiwanese Language in the Process of Making Multi-Linguistic Policies. In proceedings of the 14th Annual Conference of European Association of Taiwan Studies. Ca'Foscari University of Venice, Italy.

Sio, J. U.-S., & Morgado da Costa, L. (2019). Building the Cantonese Wordnet. In Proceedings of the Tenth Global Wordnet Conference (pp. 206-215). Wrocław, Poland.

Sybesma, R. P. E., Behr, W., Gu, Y., Handel, Z. J., Huang, C.-T. J., & Myers, J. (Eds.). (2017). *Encyclopedia of Chinese Language and Linguistics* (Vol 4). Brill.

Wee, L.-H, & and Li, M. (2015). Modern Chinese Phonology. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 474-489). Oxford University Press.

[Xiong, K., & Tamaoka, K.] 熊可欣 - 玉岡賀津(2014）「雄日中同形二字漢字語の品詞性の対応関係に関 する考察」『ことばの科学 第 27 号 (特集号)，25-51.

[Matsushita, T., Chen, M., Wang, X., & Chen, L.] 松下達彦･陳夢夏･王雪竹･陳林柯（2017）「日中対照漢字 語データベースの開発と応用」『日本語教育会秋季大会予稿集』，336-371.

[Wang, Y., Xu, C. & Kodama, S.] 王永全 - 許昌福 - 小玉新次郎（2007）『日中同形異義語辞典』東方書店. [Tang, L.] 唐磊（1993）『現代日中常用漢字対比詞典』北京出版社.

[Teramura, H.] 寺村秀夫(1990)『外国人学習者の日本語誤用例集』（大阪大学；PDF 版、国立国語研究 所、2011 年）

A PROTOTYPE AFRIKAANS ONLINE DICTIONARY FOR ACADEMIC EDITING PURPOSES

Maret Blom
Stellenbosch University, South Africa
maretblom@sun.ac.za

Abstract

The academic writing and research skills of postgraduate students at higher education institutions in the South African context are inadequate (Van Aswegen 2007), and therefore there is a growing demand for the services of editors of specifically academic texts. The editors of Afrikaans academic texts, however, experience problems in terms of reference sources (style guides or standardisation sources) that they can use to ensure consistency in the academic texts. Consequently, these editors have a need for an Afrikaans dictionary that is specially aimed at the needs of academic editors (Blom, 2020).

A model for the design of an Afrikaans online dictionary for academic editing purposes (Blom's 2018 dictionary model) was set up as the first part of a larger project that aims to compile a complete online Afrikaans academic editing dictionary. In order to implement Blom's 2018 dictionary model to compile a complete dictionary, a prototype dictionary must be developed and then tested by its target users (i.e. academic editors).

In this article an integrated theory was used to compile the said prototype Afrikaans academic editing dictionary. Fuertes-Olivera and Tarp's (2014) function theory for specialised online dictionaries was used to determine the functions and datatypes of the academic editing dictionary. The principles of the ISO standards (ISO-standard 9241-110, ISO-standard 9241-11 and ISO-standard 9241-12) and the interaction design (Sharp, Rogers & Preece, 2007) were integrated with the lexicographical theories and used to select the technological features and usability of the online dictionary. The integrated theory enables the compilation of a user friendly prototype dictionary that will be evaluated through repeated usability testing with the end or target users as test participants in a follow-up study.

Keywords: Academic editing dictionary, electronic lexicography, usability studies, interaction design.

1 Introduction and background

Higher education institutions require that students' theses must be edited as part of the compilation and finalisation process (Law, 2011). In addition, according to Van Aswegen (2007), many of these postgraduate students at higher education institutions in the South African context have insufficient academic writing skills and a shortage of research skills, which causes them to experience problems writing their theses or research reports. The fact that higher education institutions require editing, and students have poor academic writing skills, leads to a definite demand for the service of editors who specifically edit academic texts. The Stellenbosch University (SU) Language Centre, for example, offers a text editing service where experienced language practitioners edit a wide range of texts, including lecturers' and postgraduate students' theses, dissertations and articles. According to the SU Language Centre (2019), these academic editors need to perform a multitude of tasks (see SU Language Centre's website for a desription of these tasks).

In order to perform these tasks, academic editors need to make use of tools to check the language usage and help resolve different problems that arise during the editing process (Carstens & Van de Poel 2012). According to Carstens and Van de Poel (2012) no human being, including text editors, has the complete knowledge of a language and that is why editors should strive to get the best possible support (i.e. reference sources) to solve problems in the editing process. In the Afrikaans academic editing practice it is, however, a problem that some of the reference sources in Afrikaans are outdated and some of the subject dictionaries are out of print and difficult to obtain (Carstens & Van de Poel, 2012).

In an investigation into the current Afrikaans reference sources that are available and useful to academic editors, the websites of the South African Translators' Institute (SATI), the Professional Editors' Guild (PEG) and the SU Language Centre indicated that there are many style guides and terminology lists available in Afrikaans; as well as specialised dictionaries that deal with the theoretical aspects and specialist field of text editing, and dictionaries for specific subject areas (Blom, 2020). However, these Afrikaans reference sources are fairly widely distributed, on the websites of SAVI, PEG and the SU Language Centre and none of these sources contain all the aspects that an academic editor needs to complete his/her editing service fast and accurately (Blom, 2020). In the international context, the only resources specifically devoted to the field of academic editing are largely aimed at the student who has to do academic writing. (See *Collocaid, Academic writing assistant, Write-it* and *Skryfhulp*). In Blom (2020) it was concluded that there is not yet, within the international or South African context, an online reference source that has been compiled specifically for the academic editor as a target user. There are also no research that has been done on editors of specifically Afrikaans academic texts' use of currently available reference sources and the possible need for a new Afrikaans reference source.

With the aim of compiling a one-stop Afrikaans reference source where academic editors can access all the necessary information regarding academic editing in one place, a model for the design of an Afrikaans online dictionary for academic editing purposes (Blom's 2018 dictionary model) was set up. Fuertes-Olivera and Tarp's (2014) function theory for specialised online dictionaries was used as a basis for Blom's 2018 dictionary model. Fuertes-Olivera and Tarp (2014) divide the design, compilation and updating of specialised online dictionaries into three phases, namely the pre-compilation phase, the compilation phase and the post- compilation phase. Blom's 2018 dictionary model was mainly designed according to Fuertes- Olivera and Tarp's (2014) pre-compilation phase and compilation phase so that the dictionary's functions and data types could be determined.

In the pre-compilation phase decisions were made about the potential users in a specific situation and the lexicographically relevant information needs that these users may have in the specific situation (Fuertes-Olivera & Tarp, 2014). In table 1 the user profile for the academic editing dictionary are summarized according to the characteristics of the advanced academic editors, as well as the students who are still receiving training to practice as editors. The needs of the academic editors and the situations and purpose for which the editors of Afrikaans academic texts will use an online Afrikaans academic editing dictionary are also summarized in table 1.

Table 1: User Profile for the academic editing dictionary (Blom, 2020)

Target users and characteristics	1. Advanced academic editors and second year, third year and postgraduate students (semi-experts/ laymen2 in relation to the different subject areas).
User needs	1. **Primary user needs:** Academic editors have a need for: - technical aspects (academic reference systems; format of different texts; table of contents; division and numbering of chapters; format of tables, lists and graphs) - language, spelling and punctuation rules (capitalisation, sentence construction, hyphens)

	Other information such as: - subject terms from different subject areas - abbreviations and acronyms - description of the editor's role - proper names frequently used in academia - detailed sample material applicable to academic Afrikaans 2. **Secondary user needs:** A user guide that shows the editors how to use the online dictionary.
Usage situations and corresponding functions	**1. Communicative situations** **-** Text reception, text production, text correction **2. Cognitive situations** - Information on the specific subject area in which the text was written, for example chemistry. **3 . Operative situations** - Editing guidelines that give instructions as to the extent to which it is ethically acceptable to edit an academic text. - Information on the specialist field of academic editing.

The first task of the compilation phase is to compile a lexicographic database to store the data that is useful for the academic editors. The usage situations (communicative, cognitive and operative situations) and the corresponding functions (text production, text reception, text correction, additional information on specific subject areas and guidelines on ethical editing) in table 1 were used in table 2 to give an outline of the data types that could possibly be included in a lexicographic database for the academic editing dictionary (Blom, 2020).

In table 2 the various data types are listed in the left column and the reason why this data type can be found in the academic editing dictionary are listed in the right column. A typical entry in the academic editing dictionary will contain a lemma and a meaning paraphrase of the lemma, as well as, where applicable, cotext information such as additional grammatical data or collocations of the lemma that can help the editor use or understand the lemma in an academic text (Blom, 2020). Furthermore, it is important that cotext entries such as example sentences are provided, so that the editor can see how to use the lemma within the sentence context of an academic text. The lexicographic and proscriptive notes also give the editor more clarity on the correct use of the lemma in, for example, a specific field of study. If the specific entry of the lemma is not sufficient, the editor can always use hyperlinks to go to a cross-reference to another entry in the academic editing dictionary, or an external source, such as another online dictionary or web page.

Table 2: Data types in the lexicographical database for the Academic Editing Dictionary (Blom, 2020)

Data type	**Rationale**
Lemma	Most dictionaries describe lemmas.
Grammatical data	Inflections of the lemma to assist the editors with text production, as well as rules on correct sentence constructions.
Meaning paraphrase	The meaning(s) of the lemma to assist the editors incommunicative situations: text production, text correction and possibly text reception function.
Collocations	Short and long phrases to help the editors in communicative situations: text production and text correction.

Examples	Full sentences indicating the lemma in use to assist the editors in cognitive and communicative situations: text production and text correction.
Extra source	Hyperlinks to external texts to assist editors in cognitive, operational and communicative situations: text production, text reception and text correction.
Lexicographical notes	Usage notes addressed to the lemma and indicating the specific usage and cultural details to assist the editors in cognitive and operative situations.
Proscriptive notes	Notes used for recommending specific lemmas to assist the editors in communicative situations: text production, text correction, and possibly text reception. The notes can also be presented as ethical guidelines that give the editor in operative situations guidelines on how much he/she may edit on a specific issue.
Cross-references	Hyperlinks to internal texts to assist the editors in cognitive and communicative situations: text production, text reception and text correction.

The user profile of the academic editor in table 1 and the data that the academic editing dictionary should contain in table 2 were used as guidelines to perform the compilation phase of the model for an online Afrikaans academic editing dictionary. Blom's 2018 dictionary model consist out of a homepage with different sections that the academic editor can use, five articles of examples from different disciplines, as well as a "mini" user guide (see Blom, 2020 for a detailed illustration of Blom's 2018 dictionary model).

The design of this model was the first part of a larger project that aims to compile a complete online Afrikaans academic editing dictionary. In the second part of this project, the academic editing dictionary must be further compiled by making decisions about the possible dictionary structures and technological features that the dictionary should display. Fuertes-Olivera and Tarp's (2014) post-compilation phase should be carried out so that the researcher can determine whether the users are satisfied with the preliminary dictionary design.

2 Problem statement, objectives, rationale and research questions

The problem that can be deduced from the background information is that the compilation of the Afrikaans academic editing dictionary has not yet been completed as the appropriate dictionary structures and technological features has not been selected and implemented. The post-compilation phase where the dictionary needs to be empirically tested has also not been carried out. In this study, the compilation phase of the academic editing dictionary is further performed to determine the dictionary structures and technological features. Blom's 2018 dictionary model and the relevant structures and features are used to compile a prototype online Afrikaans academic editing dictionary so that the usability testing of this prototype dictionary can take place in a further study.

The motivation to further compile the Afrikaans academic editing dictionary is based on Gouw's (2018) opinion that the functions, content and structures can be regarded as three integral parts of dictionaries that must also form the basis of online dictionaries. The theoretical principles of the ISO standards and the interaction design are integrated with the lexicographic theories and used to design the data presentation (layout) of the prototype, as the function and general lexicography theory does not pay attention to the technological properties and usefulness of electronic reference sources (Du Plessis, 2015).

The following two research questions are investigated to determine the data presentation and data description of the prototype academic editing dictionary:

Research question 1:

What are the academic editors' expectations of the online Afrikaans academic editing dictionary?

Research question 2:

Which principles of the ISO standards and the interaction design can be integrated and used to compile the prototype online Afrikaans academic editing dictionary?

3 Method

Research on the academic editors' expectations of the online Afrikaans academic editing dictionary and the appropriate theories for the compilation of such a prototype were carried out in 2019 and 2020 within the framework of an exploratory research design. This research design was followed to gain, in addition to the academic editor's user profile (Table 1), more insight into specifically their expectations of an academic editing dictionary, so that a prototype can be compiled based on these needs and expectations. According to Singh (2007) exploratory research is the initial research, which forms the basis of more conclusive research and it is effective in laying the groundwork that will lead to future studies. In this study, an exploratory research design is particularly useful for compiling a prototype academic editing dictionary that will form the groundwork for a usability test of this dictionary in a follow-up study. Rubin and Chisnell (2008) also emphasise that this type of early analysis and research on a product is very important, as the design decisions made in this phase will also influence later decisions about the product. It is necessary to compile the prototype dictionary according to the needs and expectations of the academic editors, as well as appropriate lexicographical and usability theoretical principles on the functions, structures and data types of the academic editing dictionary in order to create a good foundation for further decisions on the design of this dictionary.

The two research questions, as mentioned in section 2, was performed as follows to obtain the necessary information for the compilation of the prototype academic editing dictionary:

Question 1: What are the academic editors' expectations of the online Afrikaans academic editing dictionary?

The first research question was answered by conducting a qualitative investigation of the editors' expectations of an Afrikaans academic editing dictionary. A pre-test questionnaire was sent out to third year Afrikaans and Dutch 318 students[1] who developed academic editing skills in their Afrikaans Translation and Editing module. The pre-test questionnaire was also sent out to students who did Honours in Translation and took Editing Methodology and Practice as a compulsory module. According to Rubin and Chisnell (2008), a pre-test questionnaire is used to determine a user's first impressions and attitude towards a product, and therefore the researcher decided to send a pre-test questionnaire to third year and honours students to determine their attitudes and first impressions of the data presentation and data description in the prototype academic editing dictionary. This "pre test" data can also help the researcher in the follow-up usability testing of the prototype dictionary to explain the respondents' behavior, when he/she struggles to arrive at the relevant data in the article due to, for example the layout of a dictionary article (Rubin & Chisnell, 2008).

Qualitative preference data were collected through electronic data collection. The researcher collected seven respondents' opinions and feelings through the pre-test questionnaire completed on Microsoft Word documents. Ethical clearance was first applied for as the ethical guidelines of Stellenbosch University (SU) require that all research in which people are used as participants must first be submitted to the Departmental Ethics Screening Committee (DESC). Furthermore, if the participants are affiliated with

[1] The user profile of the academic editing dictionary was documented in Blom (2018; 2020) and divided into different categories according to the user's experience as an academic editor. The user profile of the academic editing dictionary consists of professional academic editors as well as students doing academic editing. Since the professional editors in Blom (2018; 2020) were used as respondents, it was decided to use a convenience sample and approach students that are available to the researcher.

Stellenbosch University, it is also required that an application for institutional clearance be made. As third year and honours students of SU were used in this study, the researcher applied at DESC and the SU Division for Information Management for institutional clearance and the application for ethical clearance was approved.

The pre-test questionnaire consisted mainly of closed questions. Rubin and Chisnell (2008) suggest in the case of closed questions that the answers for each question, position and rating should be summarised so that the researcher can see how many respondents selected each possible choice. Since only seven respondents completed the pretest questionnaire, it was not necessary to calculate the mean scores for each question and instead, the researcher focused on each respondent's answer and the reasons for their respective answers.[1] The user profile of the academic editing dictionary was documented in Blom (2018; 2020) and divided into different categories according to the user's experience as an academic editor. The user profile of the academic editing dictionary consists of professional academic editors as well as students doing academic editing. Since the professional editors in Blom (2018; 2020) were used as respondents, it was decided to use a convenience sample and approach students that are available to the researcher.

Question 2: Which principles of the ISO standards and the interaction design can be integrated and used to compile the prototype online Afrikaans academic editing dictionary?

After data collection in the first research question and the application of the principles of Fuertes-Olivera and Tarp's (2014) function theory in Blom (2018; 2020), the general lexicography theory (see Blom, 2021) and usability theories were examined to determine which principles are appropriate for compiling the prototype academic editing dictionary. As suggested by other scientists (see, among others, Du Plessis, 2015; 2017), the researcher integrated information technology with lexicography theories to enhance the academic editors' experience of the technological aspects of the academic editing dictionary. In addition to the lexicography theories, the researcher decided to follow a usability approach in this study to compile the prototype academic editing dictionary according to the principles of the ISO standards and interaction design. The reason why the principles of a usability approach were used to compile the prototype academic editing dictionary was mainly based on Du Plessis' (2017) view that a usability approach uses user experience, product effectiveness and product- human interaction to largely focus on the academic editor and how effectively, efficiently and satisfactorily he/she deals with the prototype academic editing dictionary.

The content of these theories were not described using a concrete data collection tool. The researcher merely used the students' expectations of the dictionary in the first research question and the academic editor's user profile (table 1) and data types for the academic editing dictionary (table 2) to determine how the principles of usability theories can be applied to compile the prototype dictionary accordingly.

4 Results and analysis

4.1 Research question 1: Pre-test questionnaire to student editors

The pre-test questionnaire consisted out of eight questions divided into three sections. In the first section, the researcher determined the students' user characteristics and consultation experience of online dictionaries by means of general questions. Rubin and Chisnell (2008) suggest that the users' knowledge should be determined in a specific area, so that the respondents' level of expertise in this field can be used to decide on the degree of difficulty in the dictionary. As already mentioned in section 2 and 3, in addition to the pre-test questionnaire, the respondents will complete editing tests where they have to do editing tasks using the prototype academic editing dictionary in a follow-up study. These editing tasks consist of academic paragraphs from the Natural Sciences field and therefore most of the content in the prototype academic editing dictionary is also aimed at editing problems in the Natural Sciences field. The students' knowledge in the Natural Sciences field was examined as follows in the first section of the pre-test questionnaire.

Question 1: *Did you have Physical Sciences and/or Life Sciences up to grade 12?*

The first question in the pre-test questionnaire determined whether the respondents had Physical or Life Sciences up to matric, so that the researcher could determine how much extra information the respondents needed on issues in this subject area. The respondents indicated that they are mostly laymen with regard to the Natural Sciences field, as only two out of the seven respondents had both these subjects up to matric. The content in the prototype academic editing dictionary must therefore be presented as simply and comprehensively as possible.

Question 2: *Have you ever consulted an online dictionary? If yes, name the dictionaries. If not, what other online tools do you use?*

The respondents had to indicate whether they had previously consulted online dictionaries so that the researcher could determine their experience in this regard. The respondents' consultation experience of online dictionaries is reasonable and four respondents indicated that they have previously consulted sources such as *Pharos* online, *HAT* online and *WAT* online, but as three of the respondents have not yet consulted online dictionaries, it is also necessary to pay attention to the user guide in the prototype academic editing dictionary.

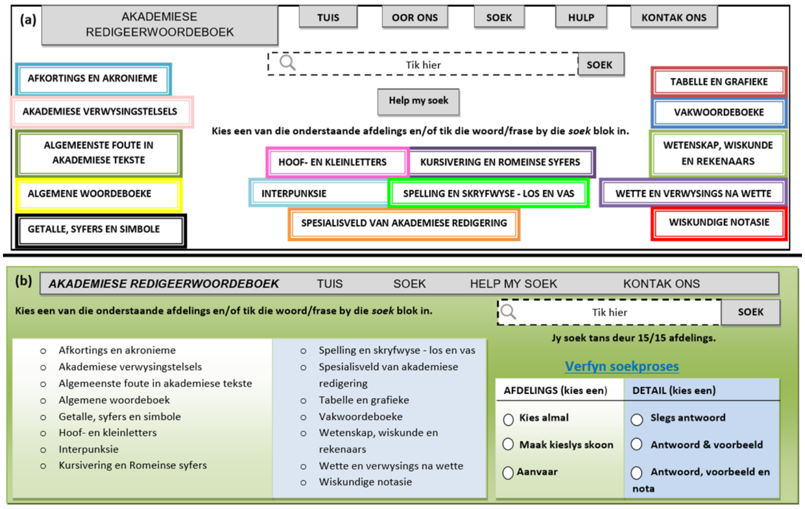
Question 3: *Mark all the unknown words in the text 1 and text 2 below (You can mark in yellow):*

In the third question the students were given two paragraphs from different academic texts where they had to mark all the unknown words. The respondents mostly marked subject terms such as, "ione", "hidronuim", "gaschromatografie-massaspektrometrie", "tetradekanoësuur" and "heksanaal". These terms could possibly be included as lemmas in the prototype academic editing dictionary, where example sentences of the term in an academic context can help the user to better understand the term.

In the second section of the pre-test questionnaire, questions were asked about the appearance of the data presentation, in other words the respondents' expectations regarding the layout of the home page, dictionary articles and user guide in the prototype academic editing dictionary.

Question 4: *Select the home page in appendix 1 in which you will be able to search for information most easily and give reason(s) for your choice:*

The students had to choose between the following four different home page layouts (see screenshots 1 and 2), including (a) a home page with extensive sections; (b) with extensive sections and a refined search process; (c) with a "Typeahead search" and a "search further" option; and (d) with broader sections:

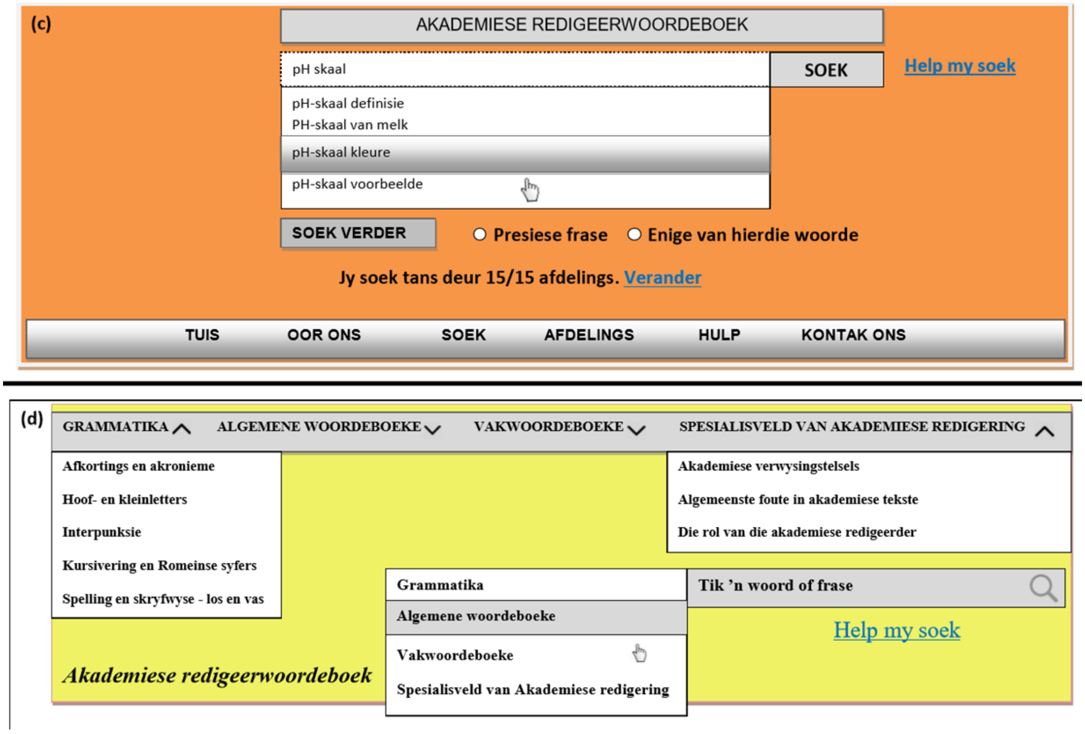

(a)
AKADEMIESE REDIGEERWOORDEBOEK
TUIS
OOR ONS
SOEK
HULP
KONTAK ONS
Tik hier
SOEK
Help my soek
AFKORTINGS EN AKRONIEME
AKADEMIESE VERWYSINGSTELSELS
ALGEMEENSTE FOUTE IN AKADEMIESE TEKSTE
ALGEMENE WOORDEBOEKE
GETALLE, SYFERS EN SIMBOLE
HOOF- EN KLEINLETTERS
KURSIVERING EN ROMEINSE SYFERS
INTERPUNKSIE
SPELLING EN SKRYFWYSE - LOS EN VAS
SPESIALISVELD VAN AKADEMIESE REDIGERING
TABELLE EN GRAFIEKE
VAKWOORDEBOEKE
WETENSKAP, WISKUNDE EN REKENAARS
WETTE EN VERWYSINGS NA WETTE
WISKUNDIGE NOTASIE
(b)
AKADEMIESE REDIGEERWOORDEBOEK
TUIS
SOEK
HELP MY SOEK
KONTAK ONS
Tik hier
SOEK
Afkortings en akronieme
Akademiese verwysingstelsels
Algemeenste foute in akademiese tekste
Algemene woordeboek
Getalle, syfers en simbole
Hoof- en kleinletters
Interpunksie
Kursivering en Romeinse syfers
Spelling en skryfwyse - los en vas
Spesialisveld van akademiese redigering
Tabelle en grafieke
Vakwoordeboeke
Wetenskap, wiskunde en rekenaars
Wette en verwysings na wette
Wiskundige notasie
Verfyn soekproses
AFDELINGS (kies een)
DETAIL (kies een)
○ Kies almal
○ Maak kieslys skoon
○ Aanvaar
○ Slegs antwoord
○ Antwoord & voorbeeld
○ Antwoord, voorbeeld en nota

Kies een van die onderstaande afdelings en/of tik die woord/frase by die *soek* blok in.

Kies een van die onderstaande afdelings en/of tik die woord/frase by die *soek* blok in.

Jy soek tans deur 15/15 afdelings.

Screenshot 1: Homepage (a) and (b) from the pre-test questionnaire

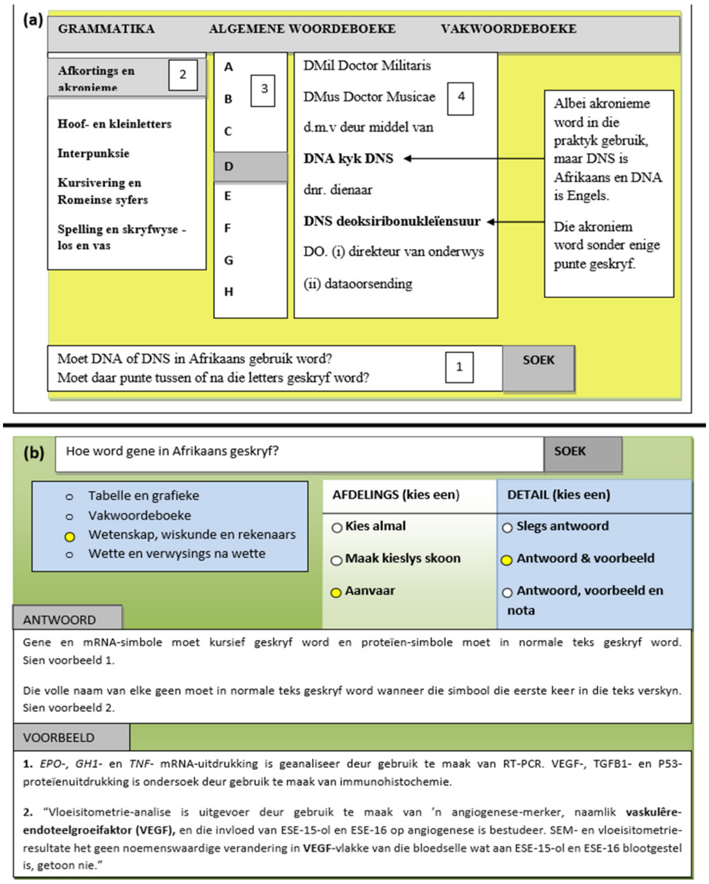

(c)
AKADEMIESE REDIGEERWOORDEBOEK
pH skaal
SOEK
Help my soek
pH-skaal definisie
PH-skaal van melk
pH-skaal kleure
pH-skaal voorbeelde
SOEK VERDER
○ Presiese frase
○ Enige van hierdie woorde
Jy soek tans deur 15/15 afdelings. Verander
TUIS
OOR ONS
SOEK
AFDELINGS
HULP
KONTAK ONS
(d)
GRAMMATIKA
ALGEMENE WOORDEBOEKE
VAKWOORDEBOEKE
SPESIALISVELD VAN AKADEMIESE REDIGERING
Afkortings en akronieme
Hoof- en kleinletters
Interpunksie
Kursivering en Romeinse syfers
Spelling en skryfwyse - los en vas
Akademiese verwysingstelsels
Algemeenste foute in akademiese tekste
Die rol van die akademiese redigeerder
Grammatika
Algemene woordeboeke
Vakwoordeboeke
Spesialisveld van Akademiese redigering
Tik 'n woord of frase
Help my soek
Akademiese redigeerwoordeboek

Screenshot 2: Homepage (c) and (d) from the pre-test questionnaire

Four respondents chose homepage (c) with a "Typeahead search" and a "search further" option and one of the explanations for this is that the respondent has no subject knowledge, therefore making it easier to type the unknown word in a search box to obtain the information. Two of the respondents chose home page (b) with extensive sections and a refined search process. The respondents' motivation is that there are more options to choose from and the fact that you can refine the sections/information will make it easier for you to find what you are looking for. One respondent chose homepage (d) with broader sections and argued that the sections are broad enough to allow for multiple definitions. The search block and main menu of the home page of the prototype academic editing dictionary can be set up according to home page (c) in screenshot 2 to make it, among other things, an easier search process for the user with a lack of subject knowledge. According to home page (b) in screenshot 3, 16 different sections can also be included on the home page of the prototype academic editing dictionary.

Question 5: *In appendix 2, select the version of results whose layout will give you the most accessible answer and give reason(s) for your choice:*

The following three versions of results that the students had to chose between, included (a) step-by-step results; (b) results following a data filter and (c) complete results with the first search process:



(a)

GRAMMATIKA ALGEMENE WOORDEBOEKE VAKWOORDEBOEKE

Afkortings en akronieme 2

Hoof- en kleinletters

Interpunksie

Kursivering en Romeinse syfers

Spelling en skryfwyse - los en vas

A B 3 C D E F G H

DMil Doctor Militaris

DMus Doctor Musicae 4

d.m.v deur middel van

DNA kyk DNS

dnr. dienaar

DNS deoksiribonukleïensuur

DO. (i) direkteur van onderwys

(ii) dataoorsending

Albei akronieme word in die praktyk gebruik, maar DNS is Afrikaans en DNA is Engels.

Die akroniem word sonder enige punte geskryf.

Moet DNA of DNS in Afrikaans gebruik word?
Moet daar punte tussen of na die letters geskryf word? 1

SOEK

(b)

Hoe word gene in Afrikaans geskryf? SOEK

- ○ Tabelle en grafieke
- ○ Vakwoordeboeke
- ● Wetenskap, wiskunde en rekenaars
- ○ Wette en verwysings na wette

AFDELINGS (kies een)	DETAIL (kies een)
○ Kies almal	○ Slegs antwoord
○ Maak kieslys skoon	● Antwoord & voorbeeld
● Aanvaar	○ Antwoord, voorbeeld en nota

ANTWOORD

Gene en mRNA-simbole moet kursief geskryf word en proteïen-simbole moet in normale teks geskryf word. Sien voorbeeld 1.

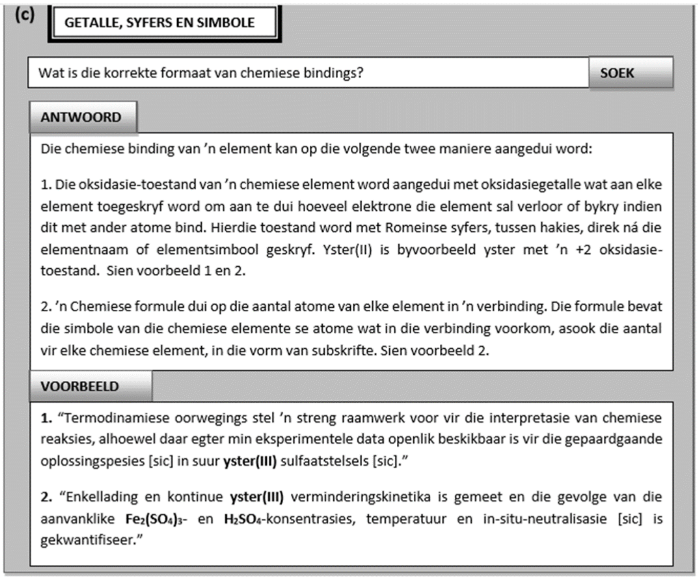
Die volle naam van elke geen moet in normale teks geskryf word wanneer die simbool die eerste keer in die teks verskyn. Sien voorbeeld 2.

VOORBEELD

1. *EPO-*, *GH1-* en *TNF-* mRNA-uitdrukking is geanaliseer deur gebruik te maak van RT-PCR. VEGF-, TGFB1- en P53-proteïenuitdrukking is ondersoek deur gebruik te maak van immunohistochemie.

2. "Vloeisitometrie-analise is uitgevoer deur gebruik te maak van 'n angiogenese-merker, naamlik **vaskulêre-endoteelgroeifaktor (VEGF),** en die invloed van ESE-15-ol en ESE-16 op angiogenese is bestudeer. SEM- en vloeisitometrie-resultate het geen noemenswaardige verandering in **VEGF**-vlakke van die bloedselle wat aan ESE-15-ol en ESE-16 blootgestel is, getoon nie."

Screenshot 3: Results (a) and (b) from the pre-test questionnaire



(c)

GETALLE, SYFERS EN SIMBOLE

Wat is die korrekte formaat van chemiese bindings? SOEK

ANTWOORD

Die chemiese binding van 'n element kan op die volgende twee maniere aangedui word:

1. Die oksidasie-toestand van 'n chemiese element word aangedui met oksidasiegetalle wat aan elke element toegeskryf word om aan te dui hoeveel elektrone die element sal verloor of bykry indien dit met ander atome bind. Hierdie toestand word met Romeinse syfers, tussen hakies, direk ná die elementnaam of elementsimbool geskryf. Yster(II) is byvoorbeeld yster met 'n +2 oksidasie-toestand. Sien voorbeeld 1 en 2.

2. 'n Chemiese formule dui op die aantal atome van elke element in 'n verbinding. Die formule bevat die simbole van die chemiese elemente se atome wat in die verbinding voorkom, asook die aantal vir elke chemiese element, in die vorm van subskrifte. Sien voorbeeld 2.

VOORBEELD

1. "Termodinamiese oorwegings stel 'n streng raamwerk voor vir die interpretasie van chemiese reaksies, alhoewel daar egter min eksperimentele data openlik beskikbaar is vir die gepaardgaande oplossingspesies [sic] in suur **yster(III)** sulfaatstelsels [sic]."

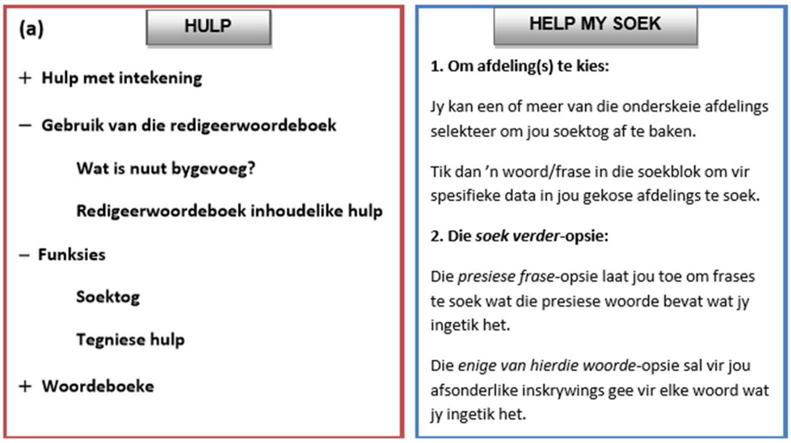
2. "Enkellading en kontinue **yster(III)** verminderingskinetika is gemeet en die gevolge van die aanvanklike **$Fe_2(SO_4)_3$**- en **H_2SO_4**-konsentrasies, temperatuur en in-situ-neutralisasie [sic] is gekwantifiseer."

Screenshot 4: Results (c) from the pre-test questionnaire

Four respondents chose (c) "the complete results with the first search process" and one of the reasons given for this is that the respondent would like the complete results with the first search as if it were a printed dictionary. The other three respondents chose (b) "the results following a data filter" and one of the reasons for this option is that the respondent can indicate how much information he/she needs, which also prevents you from reading through too much information to arrive at your answer. The layout of the results in the prototype academic editing dictionary should be similar to the layout in screenshot 4 so that users will get as much useful information as possible with their first search.

Question 6: *Choose the kind of extra help in appendix 3 that will be useful to you during the look up process and will inform you on how to look up an article in the dictionary. Give reason(s) for your choice(s):*

The students had to chose between (a) a toolbar with a navigation structure; (b) a complete user guide; and (c) help with predetermined questions:



+ Hulp met intekening
– Gebruik van die redigeerwoordeboek
    Wat is nuut bygevoeg?
    Redigeerwoordeboek inhoudelike hulp
– Funksies
    Soektog
    Tegniese hulp
+ Woordeboeke

1. Om afdeling(s) te kies:

Jy kan een of meer van die onderskeie afdelings selekteer om jou soektog af te baken.

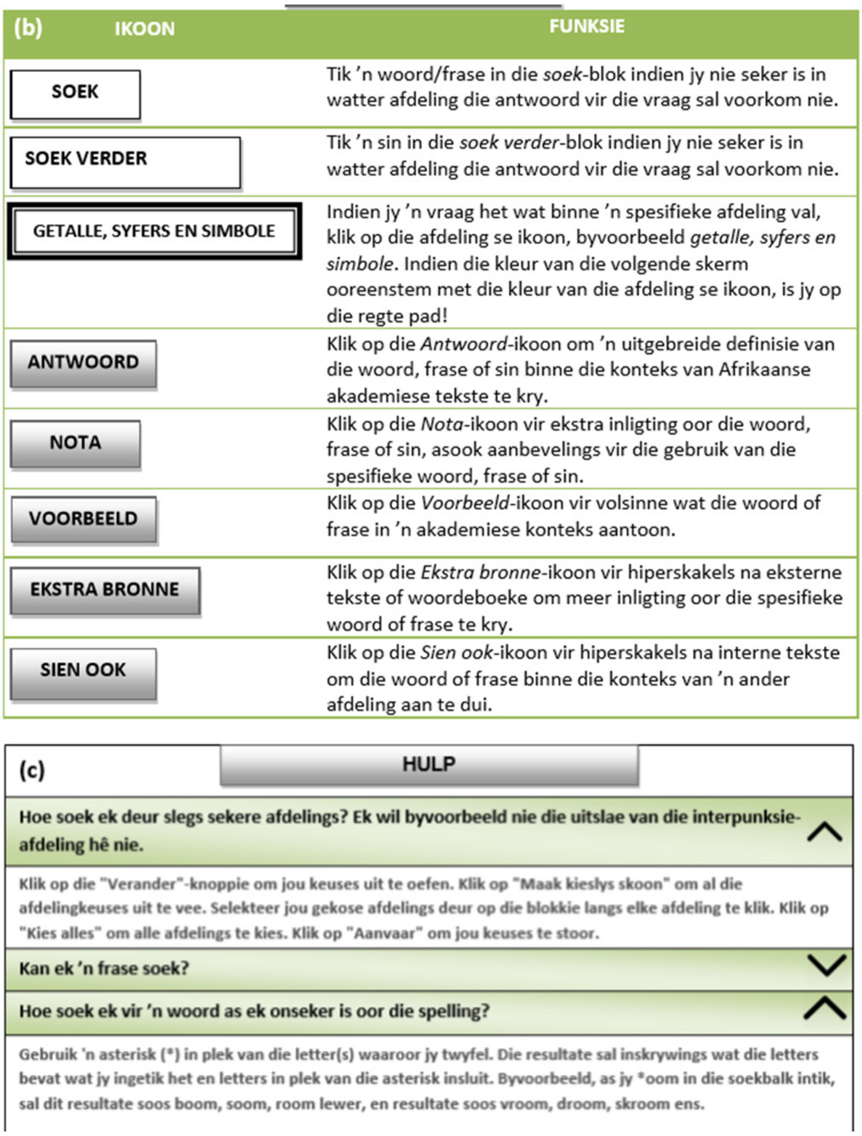
Tik dan 'n woord/frase in die soekblok om vir spesifieke data in jou gekose afdelings te soek.

2. Die *soek verder*-opsie:

Die *presiese frase*-opsie laat jou toe om frases te soek wat die presiese woorde bevat wat jy ingetik het.

Die *enige van hierdie woorde*-opsie sal vir jou afsonderlike inskrywings gee vir elke woord wat jy ingetik het.

Screenshot 5: Extra help (a) from the pre-test questionnaire



(b) IKOON	FUNKSIE
SOEK	Tik 'n woord/frase in die *soek*-blok indien jy nie seker is in watter afdeling die antwoord vir die vraag sal voorkom nie.
SOEK VERDER	Tik 'n sin in die *soek verder*-blok indien jy nie seker is in watter afdeling die antwoord vir die vraag sal voorkom nie.
GETALLE, SYFERS EN SIMBOLE	Indien jy 'n vraag het wat binne 'n spesifieke afdeling val, klik op die afdeling se ikoon, byvoorbeeld *getalle, syfers en simbole*. Indien die kleur van die volgende skerm ooreenstem met die kleur van die afdeling se ikoon, is jy op die regte pad!
ANTWOORD	Klik op die *Antwoord*-ikoon om 'n uitgebreide definisie van die woord, frase of sin binne die konteks van Afrikaanse akademiese tekste te kry.
NOTA	Klik op die *Nota*-ikoon vir ekstra inligting oor die woord, frase of sin, asook aanbevelings vir die gebruik van die spesifieke woord, frase of sin.
VOORBEELD	Klik op die *Voorbeeld*-ikoon vir volsinne wat die woord of frase in 'n akademiese konteks aantoon.
EKSTRA BRONNE	Klik op die *Ekstra bronne*-ikoon vir hiperskakels na eksterne tekste of woordeboeke om meer inligting oor die spesifieke woord of frase te kry.
SIEN OOK	Klik op die *Sien ook*-ikoon vir hiperskakels na interne tekste om die woord of frase binne die konteks van 'n ander afdeling aan te dui.

Hoe soek ek deur slegs sekere afdelings? Ek wil byvoorbeeld nie die uitslae van die interpunksie-afdeling hê nie.

Klik op die "Verander"-knoppie om jou keuses uit te oefen. Klik op "Maak kieslys skoon" om al die afdelingkeuses uit te vee. Selekteer jou gekose afdelings deur op die blokkie langs elke afdeling te klik. Klik op "Kies alles" om alle afdelings te kies. Klik op "Aanvaar" om jou keuses te stoor.

Kan ek 'n frase soek?

Hoe soek ek vir 'n woord as ek onseker is oor die spelling?

Gebruik 'n asterisk (*) in plek van die letter(s) waaroor jy twyfel. Die resultate sal inskrywings wat die letters bevat wat jy ingetik het en letters in plek van die asterisk insluit. Byvoorbeeld, as jy *oom in die soekbalk intik, sal dit resultate soos boom, soom, room lewer, en resultate soos vroom, droom, skroom ens.

Screenshot 6: Extra help (b) and (c) from the pre-test questionnaire

Six of the respondents chose (b) "the complete user guide". The motivation is that it will be the fastest to use, as you only have to search for the right icon and not answer an entire question. It is also very user friendly, especially if you have never used online dictionaries before. One of the respondents chose (a) "a toolbar with a navigation structure", and argued that user guides tend to be overly densely written, and predetermined questions do not necessarily cover all the problem areas. One of the respondents also indicated that user guide (b) in screenshot 6 might be better for trusted and regular users of online dictionaries and that additional help should be provided to the users similar to user guide (a) and (b) in screenshots 5 and 6. The extra help in the prototype academic editing dictionary must include the information given in screenshots 5 and 6 to be able to help users with different dictionary consulting capabilities.

The third section deals with the data description or content that should appear in the dictionary articles and user guide of the prototype academic editing dictionary.

Question 7: *Edit the paragraph on the next page using the data in the screenshots given in appendix 2. Which screenshot did you use and why:*

Following a paragraph from the Afrikaans summary of a Master's study on "The occurrence of Shiga-toxin producing Escherichia coli in South African game species" which the respondents had to edit, they were asked which screenshot they used and why they used that particular screenshot. Most of the respondents used both results (a) and (b) of screenshot 3, as the information there was logically set out and they understood it better than the information on the other screenshots given to them. However, the respondents also mentioned that they needed more information on the specific subject issue and that the screenshots did not completely overcome the lack of subject knowledge. The content that appears in the prototype academic editing dictionary must therefore pay sufficient attention to the subject terms of specifically the Natural Sciences field.

Question 8: *How do the explanations in the screenshots in appendix 3 help you? What additional help information would you still like?*

The respondents indicated that the extra help screenshots give them a basic idea of how the dictionary works as well as provide more information on complicated issues. However, they feel that this help can be improved by being more detailed, including examples from the dictionary as an assurance that the search process will bring them to the correct answer and provide more help to overcome the lack of subject knowledge.

4.2 Research question 2: Integration of lexicography and usability theories

Usability theories were examined and compared to lexicographical theories to determine the best way to present information in the academic editing dictionary. Tarp (2000) believes that everything in a dictionary, in other words the dictionary structure and content, is co-determined by the dictionary's functions. In this study, the same approach is followed, because the functions of the academic editing dictionary as established in Blom (2018; 2020) on the basis of Fuertes-Olivera and Tarp's (2014) function theory for specialised online dictionaries, are used as a framework to decide which principles of the usability theories are relevant for the compilation of the academic editing dictionary.

4.2.2 Usability approach and interaction design

Usability approach

When developing digital tools such as an online dictionary, the interaction between the dictionary and the computer software as well as the interaction between the user and the software are important, as these interactions determine how effectively the online dictionary can be used by the users (Du Plessis, 2017).

According to Du Plessis (2015), three chapters from ISO standard 9241 are important for creating usable electronic reference sources. (See ISO standard 9241-110: 2020 (2006); ISO standard 9241-11: 2018 (1998a) and the ISO standard 9241-12: 2017 (1998b)). According to Du Plessis (2017), these principles form the basis of any usability study and should be integrated with and applied to online dictionaries. The usability approach, which includes Du Plessis' (2015; 2017) stated set of principles of the ISO standards, is integrated in this study with the already existing lexicographic principles of the function theory and general lexicography to compile a useful prototype academic editing dictionary.

The following principles from Du Plessis (2015; 2017), which emphasise the presentation of the data and the user's experience with the software, must be taken into account when compiling the prototype academic editing dictionary:

1. *Task suitability*: The user interface of the academic editing dictionary must be suitable for processing a series of tasks (communicative, cognitive and operative tasks) and presenting the data in such a way that the academic editor can interpret it.
2. *Self-descriptions*: The user interface of the academic editing dictionary should be able to provide clear feedback, for example feedback on unsuccessful searches.
3. *Clarity and neatness*: The data in the academic editing dictionary should appear in appropriate colors, fonts and font sizes so that the academic editors can easily spot and read it.
4. *Controllability and Discriminability*: The academic editors must be in control of the interactive elements of the academic editing dictionary. For example, a main menu can help you to move to specific dictionary sections.
5. *Consistency with user expectations*: The presentation of the data in the academic editing dictionary should match the academic editors' expectations of how data should appear in online dictionaries. This data must also be presented consistently using the same article structure and display format throughout.
6. *Conciseness and traceability*: A dictionary article in the academic editing dictionary should draw the academic editor's attention only to the relevant information, so that the user is not overloaded with data or struggling to locate the relevant data.
7. *Error checking*: The academic editing dictionary should guide the academic editors in the right direction when they make mistakes, for example when a term is misspelled in the search block.
8. *Individual suitability*: The user interface of the academic editing dictionary must be able to be manipulated so that each user can create his/her own profile.
9. *Learning suitability*: The academic editing dictionary should be simple enough for all academic editors to understand. They should also be able to save, for example, previous searches and favorite searches.

According to Du Plessis (2017), usability is not only based on the mentioned ISO standards, but also on other theoretical paths such as the interaction design (Sharp, Rogers & Preece, 2007), which combine the core principles of the ISO standards and other human-computer interaction principles from the information technology. The following theoretical principles of the interaction design are directly related to usability studies and also correspond to Fuertes- Olivera and Tarp's (2014) principles for compiling specialised online dictionaries. These principles are therefore also important for the compilation process of the prototype academic editing dictionary.

Interaction design

The interaction design consists of the development and design of interactive products with the users and their individual experience as the central feature of the interaction design. According to Du Plessis (2017), the interaction design and function theory are based on the same principle, namely that the designer/lexicographer must first identify the potential users. (See, for example, Fuertes-Olivera and Tarp, 2014 for an explanation of the pre-compilation phase of an online dictionary, where decisions about the potential users must also be made.) Sharp et al. (2007) further argue that the designer/lexicographer must determine what experience these users are looking for when using the product. Sharp et al. (2007) state that the user experience of an online dictionary cannot be designed, but can take place in a framework that can be developed from four steps/activities. First, the user needs and desired user experience must be determined, secondly conceptual or physical models of the product must be designed, thirdly, prototypes must be built to further analyse the product and finally, the product must be evaluated in terms of usability principles so that the final product can be manufactured.

Due to the overlap between the interaction design and function theory, the academic editing dictionary's target users and their needs, as well as the compilation of Blom's 2018 dictionary model have already been established in the usability approach according to the first two steps of the interaction design. The target users of the academic editing dictionary have already been identified in Blom (2018; 2020) as editors of Afrikaans academic texts and the needs of these editors have also been identified as, among other things, a need for subject terms from different subject areas, language, spelling and punctuation rules and a description of the editor's role (Blom, 2018; 2020). Subsequently, Fuertes-Olivera and Tarp's (2014) second step, namely the compilation phase of the dictionary/dictionary model, was performed to compile Blom's 2018 dictionary model. Again, this phase corresponds with Sharp et al.'s (2007) second step, which requires models to be set up to satisfy user needs. The third step, namely building interactive versions of these models, is performed in this study to compile a prototype academic editing dictionary based on Blom's 2018 dictionary model. The prototype will then be evaluated in terms of usability, acceptability and effectiveness in a follow-up study during step four (Sharp et al. 2007).

Fuertes-Olivera and Tarp's (2014) post-compilation phase suggests that the lexicographers should make the dictionary available to the users, observe how it works and determine whether the users are satisfied with it. This phase, that corresponds with Sharp et al.'s (2007) last two steps, also involves continuous updating of the dictionary. As already mentioned, the post- compilation phase of the preliminary academic editing dictionary is performed in this study to determine whether the academic editors are satisfied with the academic editing dictionary in a follow-up study. It is precisely in the post-composition phase where there is a gap in the function theory, since, as Du Plessis (2015) mentions, no reference is made to the work done in the usability approach. The theoretical principles of the usability approach and interaction design are therefore used in the post-compilation phase to compile the prototype academic editing dictionary. Thereafter usability tests are performed in which the prototype academic editing dictionary is given to prototypical academic editors to test.

5 Prototype Afrikaans academic editing dictionary

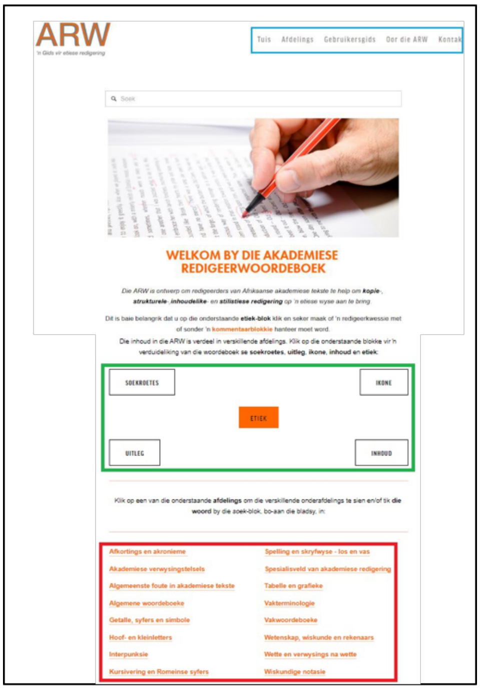
The academic editing dictionary is a prototype for an online dictionary that contains the following 16 different sections: data on abbreviations and acronyms; academic reference systems; most common errors in academic texts; general dictionaries; numbers and symbols; uppercase and lowercase letters; punctuation; italics and Roman numerals; spelling; the specialist field of academic editing; tables and graphs; subject terminology; subject dictionaries; science, mathematics and computers; laws and references to laws and mathematical notation. The purpose of the academic editing dictionary is to provide information for editors who edit Afrikaans academic texts such as theses, dissertations and scientific articles.

The prototype academic editing dictionary is developed to test the functionality of the dictionary's data presentation and data description in a follow-up study. The design of the prototype academic editing dictionary is, as already mentioned, based on lexicography and usability theories and was compiled in Blom (2021) to appear online with the help of a website builder called Squarespace. According to the website Builder Expert (2020), Squarespace is designed to help people build their own websites and then display their work, regardless of their technical skills or coding knowledge. Although this website builder is primarily aimed at building websites and cannot necessarily be used to compile a complete online dictionary, this program, taking into account cost, time and the researcher's lack of coding knowledge, is sufficient to compile a simple prototype academic editing dictionary.

The data presentation/layout of the prototype

In screenshot 7, the home page of the prototype academic editing dictionary is displayed with a search block, icons that guide the academic editor to explanations of how the prototype's search routes, layout, icons, content and ethics should be followed and interpreted (see block markedin green). There are also links to the 16 different sections from which the academic editors can extract information (circled in red). The data presentation or layout of the prototype's homepage is compiled in accordance with the principles from Du Plessis (2015; 2017) which emphasises the presentation of the data and the user's experience with the software. The results of the pre-test questionnaire are also taken into account.

First, attention is paid to the *clarity and neatness* (Du Plessis 2015; 2017) of the data on the home page. The white background in screenshot 7 makes it easier for the users to read the black and orange letters, while the uppercase and lowercase letters, bold and italics highlight important data for the users. Secondly, the main menu in screenshot 7 (circled in blue) and the 16 interactive sections help the users with *controllability and discriminability* (Du Plessis 2015; 2017) so that they can move to different parts in the prototype academic editing dictionary. Thirdly, the prototype academic editing dictionary's home page in screenshot 7 has been set up in *accordance with user expectations* (Du Plessis 2015; 2017). In the pre-test questionnaire (see section 4.1), the respondents indicated that they prefer the layout of home page (c) in screenshot 2 and that they choose homepage (b) in screenshot 1 as a second option. The search block, main menu and colour scheme of the prototype academic editing dictionary's home page in screenshot 7 are therefore compiled according to screenshot 2 and the 16 different sections, as displayed in screenshot 1 are also included on the home page of the prototype academic editing dictionary.



WELKOM BY DIE AKADEMIESE REDIGEERWOORDEBOEK

Die ARW is ontwerp om redigeerders van Afrikaanse akademiese tekste te help om ***kopie****-,* ***strukturele****-,* ***inhoudelike****- en* ***stilistiese redigering*** *op 'n etiese wyse aan te bring.*

Dit is baie belangrik dat u op die onderstaande **etiek-blok** klik en seker maak of 'n redigeerkwessie met of sonder 'n **kommentaarblokkie** hanteer moet word.

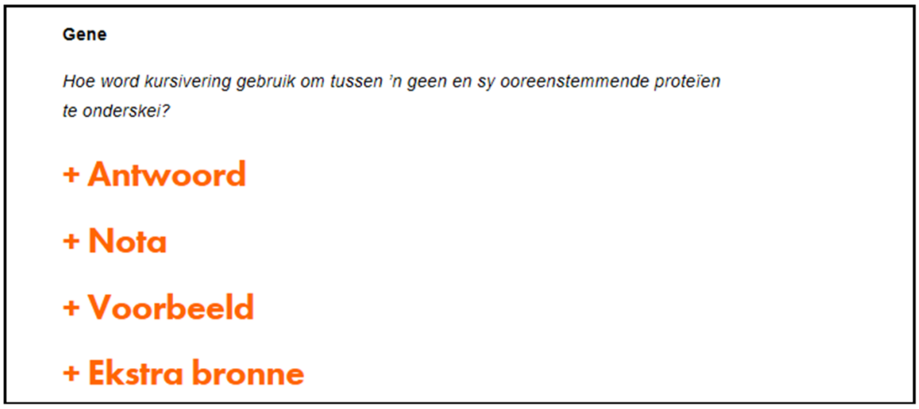
Die inhoud in die ARW is verdeel in verskillende afdelings. Klik op die onderstaande blokke vir 'n verduideliking van die woordeboek se **soekroetes**, **uitleg**, **ikone**, **inhoud** en **etiek**:

Klik op een van die onderstaande **afdelings** om die verskillende onderafdelings te sien en/of tik **die woord** by die soek-blok, bo-aan die bladsy, in:

Afkortings en akronieme
Akademiese verwysingstelsels
Algemeenste foute in akademiese tekste
Algemene woordeboeke
Getalle, syfers en simbole
Hoof- en kleinletters
Interpunksie
Kursivering en Romeinse syfers
Spelling en skryfwyse - los en vas
Spesialisveld van akademiese redigering
Tabelle en grafieke
Vakterminologie
Vakwoordeboeke
Wetenskap, wiskunde en rekenaars
Wette en verwysings na wette
Wiskundige notasie

Screenshot 7: Home page of the prototype academic editing dictionary

The data layout of the dictionary articles in the prototype academic editing dictionary is also compiled according to Du Plessis' (2015; 2017) discussion of the principles of the ISO standards (9241-110; 9241-11; 9241-12) and the results of the pre-test questionnaire. After completing the pre-test questionnaire, most respondents indicated that they preferred the data layout in screenshot 4, as it is simple enough and clearly set out. In screenshots 8 and 9, the dictionary articles for the lemmas "genes" and "chemical bonds" in the prototype academic editing dictionary are compiled in *accordance with user expectations* (Du Plessis 2015; 2017), as the same article structure as in screenshot 4 is used. Furthermore, the data in screenshots 8 and 9 were also presented consistently using the same article structure and display format in both screenshots. The *conciseness and traceability* (Du Plessis 2015; 2017) of the dictionary articles in the prototype academic editing dictionary is ensured by first hiding the data under each icon (see screenshot 8). For example, if the academic editor decides to click on the "voorbeeld" icon (see screenshot 9), the most important data is in bold. The academic editor is therefore not overloaded with data.



Gene

Hoe word kursivering gebruik om tussen 'n geen en sy ooreenstemmende proteïen te onderskei?

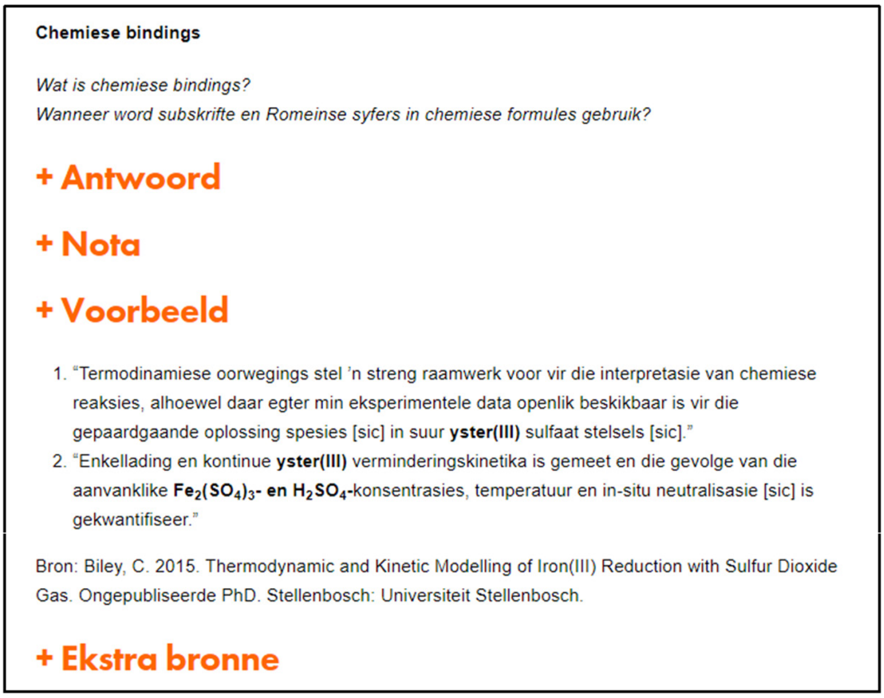
+ Antwoord

+ Nota

+ Voorbeeld

+ Ekstra bronne

Screenshot 8: Layout of the dictionary article for the lemma "genes" in the prototype



Chemiese bindings

Wat is chemiese bindings?
Wanneer word subskrifte en Romeinse syfers in chemiese formules gebruik?

+ Antwoord

+ Nota

+ Voorbeeld

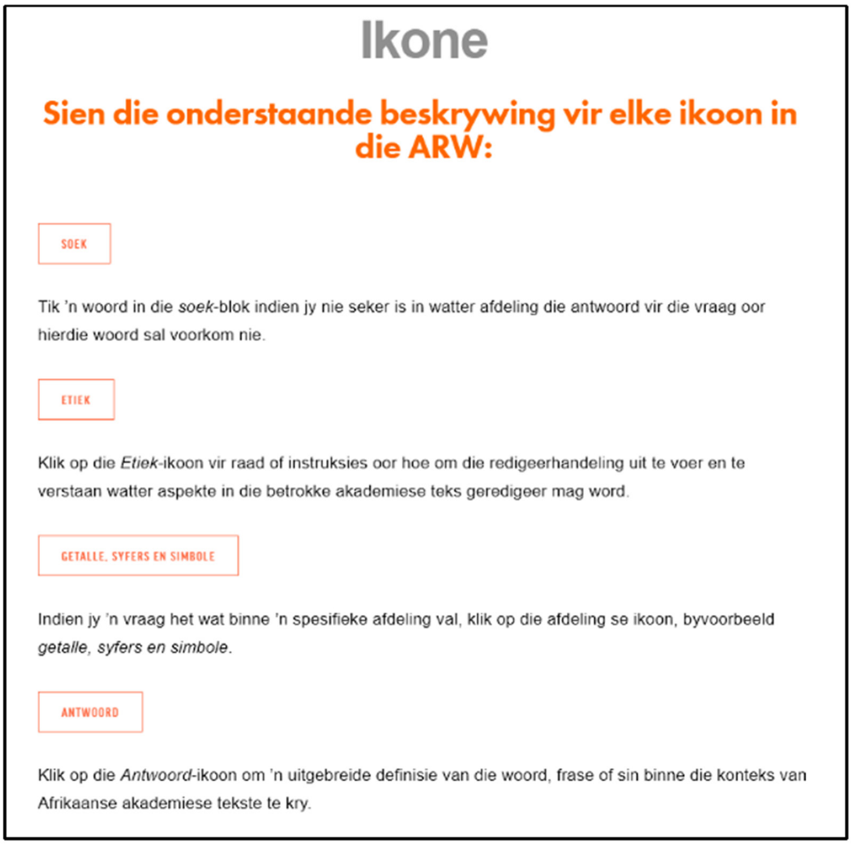
1. "Termodinamiese oorwegings stel 'n streng raamwerk voor vir die interpretasie van chemiese reaksies, alhoewel daar egter min eksperimentele data openlik beskikbaar is vir die gepaardgaande oplossing spesies [sic] in suur **yster(III)** sulfaat stelsels [sic]."
2. "Enkellading en kontinue **yster(III)** verminderingskinetika is gemeet en die gevolge van die aanvanklike **$Fe_2(SO_4)_3$- en H_2SO_4**-konsentrasies, temperatuur en in-situ neutralisasie [sic] is gekwantifiseer."

Bron: Biley, C. 2015. Thermodynamic and Kinetic Modelling of Iron(III) Reduction with Sulfur Dioxide Gas. Ongepubliseerde PhD. Stellenbosch: Universiteit Stellenbosch.

+ Ekstra bronne

Screenshot 9: Layout of the dictionary article for the lemma "chemical bonds" in the prototype

In the pre-test questionnaire, the respondents indicated that they prefer the layout of extra help (b) in screenshot 6 as they only need to search for the relevant icon and can get to the required information as quickly as possible. The layout of the "icon" user guide in the prototype academic editing dictionary has been compiled in *accordance with user expectations* (Du Plessis 2015; 2017) (see screenshot 10 for an outline of the prototype academic editing dictionary's "icon" user guide).



Ikone

Sien die onderstaande beskrywing vir elke ikoon in die ARW:

SOEK

Tik 'n woord in die *soek*-blok indien jy nie seker is in watter afdeling die antwoord vir die vraag oor hierdie woord sal voorkom nie.

ETIEK

Klik op die *Etiek*-ikoon vir raad of instruksies oor hoe om die redigeerhandeling uit te voer en te verstaan watter aspekte in die betrokke akademiese teks geredigeer mag word.

GETALLE, SYFERS EN SIMBOLE

Indien jy 'n vraag het wat binne 'n spesifieke afdeling val, klik op die afdeling se ikoon, byvoorbeeld *getalle, syfers en simbole*.

ANTWOORD

Klik op die *Antwoord*-ikoon om 'n uitgebreide definisie van die woord, frase of sin binne die konteks van Afrikaanse akademiese tekste te kry.

Screenshot 10: A partial screenshot of the icon user guide in the prototype academic editing dictionary

Data description/content of the prototype

The data description or content of the prototype academic editing dictionary was selected in accordance with the prototype's functions (communicative, cognitive and operative), dictionary structures (data layout) and results of the pre-test questionnaire. As already mentioned, the prototype is only a preliminary version of the dictionary and is specifically being compiled to test the effectiveness of the content. The content is also compiled in accordance with the editing tests that will be written during the usability testing. These editing tests consist of academic paragraphs from the Natural Sciences field and therefore most of the content in the prototype academic editing dictionary is also aimed at editing problems in the Natural Sciences field. The respondents indicated in the pre-test questionnaire that they are mostly laymen with regard to the Natural Sciences field and therefore the content in the prototype academic editing dictionary is presented as simply and comprehensively as possible. The content that appears in the prototype academic editing dictionary must pay sufficient attention to the subject terms of specifically the Natural Sciences field. Furthermore, for the sake of the text reception, text production, text correction and the operative function of the academic editing dictionary, the following aspects must also be included in the prototype: technical aspects on academic texts, language, spelling and punctuation rules, abbreviations, a description of the editor's role, proper names that are frequently used in academia and detailed sample material that is applicable to academic Afrikaans.

As already explained with the layout of the prototype's home page, the content in the prototype academic editing dictionary is divided into 16 different sections. The content of the prototype academic editing dictionary and the manner in which it is set out for the academic editors as a user guide in the prototype are given in table 3 on the next page. See the respective sections in this table for an outline of the data included in each section of the prototype academic editing dictionary.

Table 3: A breakdown of data in each section of the prototype academic editing dictionary

Abbreviations and acronyms

This section contains a list of abbreviations and acronyms that typically occur in academic texts. Each abbreviation/acronym is written out in full under the "answer" icon and in some cases a description of the abbreviation/acronym is also given. Additional information about, for example, the origin of the abbreviation/acronym is also given under the "note" icon. The "example" icon gives example sentences in which the abbreviation/acronym is used in academic texts so that the editor can see how to use the abbreviation/acronym correctly in an academic context.

Academic reference systems

For the time being, this section focuses only on the Harvard reference system, but the full dictionary will also outline other reference systems such as APA and Vancouver. The basic features of the Harvard system are given, as well as an explanation of the way in which the different types of sources, such as books, academic journals and internet sources, should be referred to in the main text and in the reference list.

Common mistakes in academic texts

The purpose of this section is to help academic editors to quickly identify errors commonly made in academic texts. This section will also help novice editors become familiar with typical problems that occur in practice.

Note: this section is not yet complete, as the source that you are currently using is only a prototype.

General dictionaries

A list of links to common dictionaries is given in this section. These include the Electronic *WAT* Online, *HAT* and *Longman* Online and *Pharos* Online. There are also links to dictionaries in languages other than Afrikaans, for example Deutsch-Englisch-Wörterbuch (German/English dictionary) and Van Dale (Dutch dictionary).

Numbers and symbols

This section focuses on the correct format in which the chemical bonding of an element must be indicated. There are also example sentences to illustrate this. It explains how to insert super/subscript and the en-dash in a Word document. Furthermore, the correct way is given in which degrees Celsius (^{o}C) and percentages are written.

Uppercase and lowercase letters

General principles regarding the use of uppercase and lowercase letters are set out in this section. A link is also given to chapter 9 of the AWS (Afrikaans word list and spell list), where there are more extensive explanations of the use of uppercase and lowercase letters.

Punctuation

The concept "punctuation" is explained in this section and then reading and writing signs are discussed in a separate subsection. The use of the hyphen, quotation mark and comma is explained under "Punctuation". Links to chapters 12 and 13 of the AWS are also provided, where there are more extensive explanations on the use of the hyphen and punctuation.

Italics and Roman numerals

The general norms for the use of italics are set out in this section. The main functions are emphasis, words or phrases of foreign origin, titles and other uses.

Note: the section on Roman numerals is not complete yet, as the source you are using now is only a prototype.

Spelling and writing

The general norms about writing words togheteher and seperately are set out in this section. There is also a spell list of words that are often misused in academic texts. (Note that this list consists of only three examples and will be supplemented using the examples provided by academic editors.)

Specialist field of academic editing

The four types of editing, namely copy, stylistic, structural and content editing, as explained by Mossop (2014), are set out in this section. It is also explained what exactly editors of academic texts may do during the editing process. (In this section a table is given, which indicates which aspects the editor should edit in each category.)

A distinction is made between electronic editing and editing on printed material and a table of proofreading marks is also given, as well as the steps that must be followed to use the "track changes" function in the MS Word program.

Tables and graphs

This section focuses more on the technical aspects of an academic text. Sample material are used to demonstrate whether tables and graphs are given captions above or underneath the table or graph.

Subject terminology

This section consists of a list of subject terms. Each term is described and then additional information and other sources (which are usually a link to a video description of the term) are given.

Note: the information in this section is not repeated in other sections such as "Science, Mathematics and Computers". For the definition of a specific term, you should consult this section.

Subject dictionaries

This section consists of online resources that provide links to subject dictionaries from various disciplines, such as the Professional Editors' Guild, Prolingua, the Suid-Afrikaanse Akademie vir Wetenskap en Kuns and VivA.

Science, math and computers

This section covers aspects that are important for academic writing in science. For example, it is explained how genes should be written, as well as what exactly the periodic table of elements is. There is a link to an interactive version of the periodic table in Afrikaans.

Note: the section on mathematics and computers is not yet complete, as the source you are currently using is only a prototype.

Laws and references to laws

This section explains how to use italics and capital letters in academic writing in the legal field. A list of abbreviations commonly used in the legal field is also given, as well as a list of links to South African laws. Furthermore, it is explained how matters and legislation should be referred to in the main text and in footnotes. As additional resources, a list of online resources is provided with links to style and writing guides and to various websites that are useful for the legal discipline.

Mathematical notation

Note: this section is not yet complete, as the source you are currently using is only a prototype.

The next step in the development process of the prototype academic editing dictionary is to conduct a usability evaluation and test whether the technologies in this prototype have been used successfully to present the data presentation/layout and data description/content of the prototype in such a way that the academic editors could use the prototype academic editing dictionary to complete editing tasks effectively.

5 References

Blom, M. (2020). ’n Teoretiese model vir die ontwerp van ’n aanlyn Afrikaanse Akademiese Redigeerwoordeboek. *Tydskrif vir Nederlands en Afrikaans*, 27(1).

Blom, M. (2021). ’n Aanlyn Afrikaanse akademiese redigeerwoordeboek: Die bruikbaarheid van ’n prototipe. MA thesis. Stellenbosch: Stellenbosch University. 1-231.

Carstens, W.A.M. & Van de Poel, K. (2012). *Teksredaksie*. Revised version. Stellenbosch: African Sun Media.

Du Plessis, A.H. (2015). ’n Analise van die selfoon-WAT: ’n Grondslag vir die verbetering van selfoonwoordeboeke. MA thesis. Stellenbosch: Stellenbosch University. 1-221.

Du Plessis, A.H. (2017). Die rol van die bruikbaarheidsbenadering binne die e-leksikografie. *Literator*, 38(2):1-11.

Fuertes-Olivera, P.A. & Tarp, S. (2014). *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. Berlyn / Boston: De Gruyter.

Gouws, R.H. (2018). Internet lexicography in the 21st Century, in S. Engelberg, H. Kämper & P. Storjohann (eds.). *Wortschatz: Theorie, Empirie, Dokumentation*. Berlyn: De Gruyter. 215-236.

International Organisation of Standardisation. (2017). (1998b). ISO 9241-12. [Internet]. Available: https://www.iso.org/standard/64839.html [20 Oktober 2020].

International Organisation of Standardisation. (2018). (1998a). ISO 9241-11. [Internet]. Available: https://www.iso.org/standard/63500.html [20 Oktober 2020].

International Organisation of Standardisation. (2020). (2006). ISO 9241-110. [Internet]. Anvailable : https://www.iso.org/standard/75258.html [20 Oktober 2020].

Law, M. (2011). The development of core standards for editing in South Africa. *Southern African Linguistics and Applied Language Studies*, 29(3):275-292.

Law, M. (2014). Factors influencing editorial work within the sectors of the South African editing industry. *Southern African Linguistics and Applied Language Studies*, 32(3):285-299.

Mossop, B. (2014). *Revising and editing for translators*. Third edition. Manchester: St. Jerome.

Rubin, J. & Chisnell, FD. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests*. Indianapolis, IN: John Wiley.

Sharp, H., Rogers, Y. & Preece, J. (2007). *Interaction design: Beyond human computer interaction*. Chichester: John Wiley.

Stellenbosch University Language Centre. (2019). Redigering van jou tesis of navorsingsverslag. Internet. [Available]: http://www0.sun.ac.za/taalsentrum/assets/files/TaaldiensDokumente/Tesisredigering_2019_O kt.pdf [28 Maart 2020]

Singh, K. (2007). *Quantitative Social Research Methods*. SAGE Publications

Tarp, S. (2000). Theoretical challenges to practical specialised lexicography. *Lexikos*, 10:189-208.

Van Aswegen, E.S. (2007). Postgraduate supervision: The role of the (language) editor: Sed quis custodiet ipsos custodes? *South African Journal of Higher Education*, 21(8):1139-1151.

A METHOD OF DEVELOPING AN ASR MEDICAL LEXICON BASED ON THE MAPPING OF INDONESIAN PEOPLE'S SPEECH PATTERNS

Mohammad Teduh Uliniansyah, Asril Jarin, Gunarso, Eduward Butarbutar, Agung Santosa, Elvira Nurfadhilah, Lyla Ruslana Aini, Siska Pebiana
Center for Information and Communication Technology, Agency for the Assessment, Indonesia
teduh.uliniansyah@bppt.go.id

Abstract

The pronunciation of foreign terms in our acoustic data for various dialects in Indonesia, such as Javanese, Sundanese, Batak, and Minangnese, have their own unique patterns when they are notated in a pronunciation lexicon. Based on the 2010 population census data by BPS (Badan Pusat Statistik; Statistics Indonesia), the Javanese people make up 40 percent of the total population (Statistik, B. P., 2011). In this paper, we are discussing the development of a speech corpus in order to examine the patterns of the pronunciation of foreign terms by Indonesians. It turned out that the number of Javanese speakers also made up a similar proportion in our speakers' data. We propose a lexicon development method for an ASR (automatic speech recognition) modeling for medical dictation by mapping the pronunciation patterns of foreign terms. We mapped the pronunciation patterns of medical technical terms based on the recorded data of 122 speakers with various dialects. We identified which speakers with Javanese dialect and made a custom lexicon file consisting of pronunciation data for standard Indonesian dialect and Javanese dialect. The experiment results show that the ASR model built with combined standard Indonesian dialect and Javanese dialect lexicon has better accuracy than the ASR model built with standard Indonesian dialect lexicon. We hope that the proposed method can be used to build a lexicon for an ASR model intended for a multi dialects community.

Keywords: lexicon, medical terms, Javanese, automatic speech recognition, Indonesian.

1 Introduction

The main objective of this research is to develop a medical ASR system for making medical records. There are several key data needed to be collected in advance in the development of the ASR system, namely the speech corpus, the text corpus, and the lexicon data. The speech corpus is used to produce acoustic models, the text corpus is used to produce language models, and the lexicon serves to link the acoustic level representation and the output transcript of the system. The lexicon has two tasks, namely to determine what words are known by the system, and to serve as a means for building an acoustic model for each lexical item.

Speech is the most basic, common and efficient form of communication for people to interact. So, people are also more comfortable interacting with computers via speech, rather than using keyboards and pointing devices. This can be accomplished by developing an Automatic Speech Recognition (ASR) system which allows a computer to recognize the words that a person speaks and convert them into written texts. The ASR system would support many valuable applications like dictation, command and control, embedded applications, personal assistant, spoken database querying, medical applications, office dictation devices, and automatic voice translation into foreign languages. One example of the application of the ASR system in the medical field is dictation by doctors in documenting medical reports. This is in line with the research we are currently working on.

Vogel et al. (Vohel et al., 2015) conducted an evaluation of a web-based ASR system in a university hospital for medical documentation in German. They found that medical documentation with ASR increased the speed and amount of documentation, and also had a positive impact on the participants' moods compared to typing alone. Chiu et al. (Chiu et al., 2017) explored the use of ASR for medical transcription of a doctor-patient conversation. They used 14,000 hours of talk and demonstrated that the proposed model achieves promising results on important medical speech and therefore can be used practically in clinical settings for transcribing medical conversations.

Hoyt et al. (Hoyt R., 2010) evaluated speech recognition efficiency to document outpatient encounters in the EHR system at the military hospital and its 12 remote clinics. Seventy-five physicians participated to evaluate speech recognition for clinical documentation. Among these participants, 69% of doctors continue to use speech recognition in their routine practice and report that voice recognition for medical documentation is more convenient, accurate, and fast. However, the efficiency of ASR for medical documentation still needs to be improved to avoid errors that could potentially cause clinical harm. A research conducted by Goss et al. (Goss et al., 2019) showed that the accuracy of speech recognition in the medical field varies between 78% and 92%. This is affected by many things, one of which is how to pronounce medical terms that are difficult to recognize by the ASR system.

Globally, most of the medical terms come from Ancient Greek and Latin. With the advancement of health and medical science in the western world, there have been many additional medical terms derived from French, German, and Angelo Saxon (Ajami, S., 2016).

In the Indonesian language, many loanwords are adopted from Portuguese, Dutch, Arabic, etc. (Kamajaya, I et al., 2017). However, the diversity of Indonesians' pronunciations is strongly influenced by their mother tongue. As it is known, there are around 700 languages in Indonesia (Laboratorium Kebinekaan Bahasa dan Sastra, 2021), resulting in various pronunciations.

The Indonesian language has the following characteristics (Alwi, H., 2019)

- It uses the Roman alphabet, read from left to right, and the words in a sentence are separated by spaces.
- The Indonesian language has 32 phonemes: 4 diphthong phonemes, 22 consonant phonemes, and 6 vowel phonemes
- Unlike Thai, Chinese, or Vietnamese which are tonal languages, Indonesian is a non- tonal language
- The Indonesian language is slightly defective phonemic orthography language
- There are no verb conjugations. The time reference of a verb is represented by several time adverbials such as sudah (already), telah (already), nanti (later), etc.
- The Indonesian language is a genderless language.
- The Indonesian language uses complex affixes (prefixes, infixes, and suffixes) to create derivatives

2 Method

2.1 Data Collection

There are three types of data that we have prepared, namely speech corpus along with recording transcript, text corpus, and lexicon. The speech corpus is used to generate acoustic models, while the text corpus is used for creating language models. The following is a discussion of the stages in collecting the required data.

2.1.1 Speech Corpus and Transcript

Pre-recording Stage

There are two important steps to determine to obtain consistent data before the recording process takes place. The first step is preparing the transcript while the second one is selecting the speakers. The transcript was made by referencing several online medical sites and actual conversations that took place in doctors' offices through self-recording activities. The following factors were considered for selecting the speakers: number of speakers, gender composition, age range, and dialect background. We decided to involve 122 speakers with a balanced gender composition of males and females with the age range of 25 to 55 years, and various dialects such as Javanese, Sundanese, Balinese, Makassar, Papuan, Minang, Batak and Dayak.

Recording Stage

Before the recording, each speaker was given a brief explanation of the importance of consistency in pronouncing each word. In the recording process, the speakers were asked to pronounce each sentence in the transcript individually, under a phonetician supervision. Apart from that, the phonetician also guides the speakers on how to utter the sentences in a natural way.

Post Recording Stage

The final stage is the validating and editing process to ensure data accuracy, including the conformity of text and speech, noise removal, and data recapitulation.

Identification of The Speakers' Dialect

The recording process was carried out in Yogya, a city which is mostly populated by Javanese. However, there are many comers from other provinces who reside to study. The recruitment of speakers does not consider Indonesia's demographic factors. However, after identifying the speakers' dialects, the proportion of the speakers with Javanese dialect is around 48% of the total number of the speakers. This proportion is similar to the 2010 BPS census data (Statistik, B.P., 2011) It is known that the accuracy of an ASR system tailored for a particular ethnic will increase if the training data also accommodates the way the ethnic speaks. Here, we will test these phenomena by comparing the accuracy between the model with standard dialect and the accuracy of the model in which the Javanese dialect is also accommodated. For this reason, before the training process is carried out, we identify who among the speakers whose pronunciation is standard and Javanese dialect. The identification process is done by listening to how a speaker speaks. The identification results show that from 122 speakers, there are 59 speakers with Javanese dialects and 63 speakers with standard pronunciation. To enrich the resulting acoustic model, we also used the speech corpus we already had, namely the BPPT 2010 speech corpus and the BPPT 50K ASR 1 Talk Data Corpus (KDW-BPPT-50K- ASR1) (Gunawan et al., 2018).

2.1.2 Text Corpus

From several types of research on NLP (natural language processing) that we have done previously (Gunarso et al., 2016 December)(Gunarso et al., 2016)(Uliniansyah et al., 2017) (Uliniansyah et al., 2013), we collected a text corpus consisting of around 16 million unique sentences. However, the sentences are mostly not related to the medical domain. Therefore, we create a medical domain text corpus by collecting sentences from several online medical sites. Apart from that, we also enriched this medical text corpus by adding actual conversational sentences between doctors and patients. After performing a series of standard preprocessing processes such as cleansing, tokenizing, and normalizing, we obtained a medical text corpus consisting of around 3 million unique sentences comprising around 114 thousand unique words. The cleansing processes include removing HTML tags and non-ASCII characters, fixing typos, and separating sentences. The tokenizing process is separating tokens (words, characters, or subwords) within a sentence. While the normalizing process is changing symbols such as monetary units, abbreviations, and numbers to their pronunciation forms.

2.1.3 Lexicon

The lexicon has an important role in the ASR system, linking the acoustic level representation and the output transcript of the system. The lexicon has two tasks, namely to determine what words are known by the system, and to serve as a means for building an acoustic model for each lexical item. Thus, the lexicon data writing format consists of vocabulary items and pronunciation representations. On the Kaldi website page, it is written that the lexicon file format is as follows: <word> <phone1> <phone2> ... (Kaldi, 2021).

The aim of the acoustic model in ASR is to introduce phonemes in a language to the ASR system, and it is essential to ensure that the consistency of acoustic data in our speech corpus down to the phoneme level. There is a need to maintain data consistency since acoustic data in our speech corpus is collected through a recording process involving speakers with various dialects in Indonesia. Such a variety of dialects has various pronunciations of words in the transcript, which consequently induces data inconsistency.

Dealing with this cultural challenge, we need to define how words in the transcript are pronounced before the recording process takes place to obtain consistent data, and to map the patterns of how varied pronunciations occur in daily use of our speakers. The latter is to provide additional possible pronunciation to the lexicon of the system. Therefore, we apply two phonetical methods to identify possible varied pronunciations of words induced by the variety of dialects.

Defining the number of phonemes of each word roughly by manner of articulation is the first method we apply to identify whether the pronounced words have the same number of phonemes as the one in the predefined words. With this method, we can see that speakers with Javanese dialect tend to utter the typical pronunciation of the following:

- phoneme / k / is not sounded if it is in the final position as in the words ɑnɑk →ɑnɑ, *bɑtuk* → *bɑtuk*, *səsɑk* → *səsɑ*, *bəngkɑk* → *bəngkɑ*
- phoneme / ə / is not sounded if it is in between phonemes / b / and /r/, /p/ and /r/, /s/ and / l /, / s / and / r /, as in the words *bərikɑn* → *brikɑn*, *sərɑtus* → *srɑtus*, *pərɑwɑt* → *prɑwɑt*, *səlain* → *slain*

With this pattern, we can see that they tend to pronounce words with less phonemes, and this will obviously become an accuracy issue.

The second method is defining what phonemes actually make up words. After detecting that the number of phonemes pronounced by speakers matched, it is important to identify what phonemes are involved in a word. For speakers with certain dialects, they tend to pronounce the phoneme /e/ with /ə/ or otherwise. For example, speakers with Javanese dialect tend to pronounce phoneme /ɛ/ with /ə/, like in the words of *bɑktɛri* → *bɑktəri*, *tɛrɑpi* → *tə̲rɑpi*, *ɛpidɛrmis* → *ɛpidə̲rmis*, *diɑbɛtɛs* → *diɑbɛtə̲s*. In contrast, speakers with Batak dialect tend to substitute phoneme /ə/ with /ɛ/, like in the words of *bəsɑr* → *bɛsɑr*, *bənɑr* → *bɛnɑr*, *doktər* → *doktɛr*, *pərɑwɑt* → *pɛrɑwɑt*. In addition, speakers with Javanese dialect tend to sound phonemes /b/, /d/, /g/ and /ǰ/ as breathy consonants if they are in initial or before a vowel position, as in the words *bərbɑgaɪ* → *bhərbhɑghaɪ*, *bərbɛdɑ* → *bhərbhɛdhɑ*, *jugɑ* → *jhughɑ*, *bərǰɑlɑn* → *bhərǰhɑlɑn*

In other case, we can see that speakers with Bugis dialect tend to pronounce phoneme /*n*/ with /ŋ/, like in the words of *ikɑn* → *ikɑŋ̲*, *huǰɑn* → *huǰɑŋ̲*. With this pattern, we see that they tend to pronounce words with different compositions of phonemes, and this will become an accuracy issue. Therefore, to achieve accurate acoustic data in our speech corpus, we assign personnel with knowledge on phonetics to supervise the recording process.

In spite of the fact that we need to obtain consistent data in our speech corpus, we also realize that there is a potential varied pronunciation of words occur in the daily use, and this requires us to make sure that our ASR system is able to recognize both predefined pronunciation and the varied one, by providing both ways of pronunciation in our lexicon list.

2.2 ASR Modelling

The modeling process uses 3 sets of corpus, namely the 2010 BPPT corpus, the KDW-BPPT-50K-ASR1 corpus and the BPPT medical corpus with a total duration of around 245 hours.

We conducted two modeling experiments using two different lexicons: the standard lexicon and the compound lexicon. The standard lexicon contains lexical items with common Indonesian pronunciation data. The combined lexicon contains the same lexical items but the pronunciation data is adjusted according to the Javanese dialect. The Javanese lexicon dialect has its own peculiarities, especially in reading the letters b, d, g, and h, as well as for the k at the end.

The training process to produce a model is carried out using Kaldi and PyChain. Kaldi plays a role in data preparation, feature extraction, and decoding. The Kaldi output is then used by PyChain in the training process. PyChain uses PyTorch in implementing end-to-end lattice- free maximum mutual information (LF-MMI) training known as chain modeling. We use PyChain for the training process because it employs full GPU training on numerator/denominator graphs, and support for unequal length sequences. Apart from that, there are several architectural options such as TDNN, LTSM, RNN etc.

In the training process, we use the default parameters from PyChain, namely:

- TDNN architecture with 6 convolution layers
- Input dimension 40 (MFCC)
- Adam Optimizer Method
- Learning rate 0.001
- 40 Epochs

2.3 Testing the ASR Models

As previously mentioned, we have identified which speakers with standard and Javanese dialect. Some of the speakers with both dialects were separated from the whole dataset and used as test data. The evaluations were carried out using models generated from the two experiments mentioned in the previous section. The following steps were undertaken in the testing process:

- At the beginning of the testing process, a feature extraction process is carried out from the test data. The features are used as inputs for posterior generation.
- Each ASR model is used to process the previously extracted features to generate their posterior values.
- Decode the posterior values using KALDI tools to produce text transcripts.
- The text transcript results were then compared with the reference text transcripts to calculate the WER (word error rate) values.

3 Results

Table 1 shows the accuracy results and training duration of each model. Table 2 shows the accuracy results categorized according to the speakers' dialect. The common parameter for measuring the accuracy of an ASR system is WER (word error rate). It is calculated by adding up the number of insertions, deletions, and substitutions, divided by the number of words in the reference (Ali et al., 2018 July). The insertion value indicates the number of words that cannot be captured at all by the model. The deletion value indicates the total number of words that are completely wrongly captured or understood. The substitution values show the number of words that are almost correctly understood by the model (Uliniansyah et al., 2018 November).

Table 1. WER result

	% WER	**Training Duration**
Models with standard dialect lexicon	13.72	28 hours 36 minutes 38 seconds
Models with standard dialect plus Javanese dialect lexicon	13.30	29 hours 6 minutes 48 seconds

The above table shows the WER results for both models. The WER value of the model with combined dialect lexicon is slightly lower than that of the model with standard dialect lexicon. This implies that the model with the combined dialect lexicon has slightly better performance.

Table 2. WER results based on speakers' dialect

Testing data	% WER of model with standard dialect lexicon	% WER of model with standard plus Javanese dialect lexicon
Javanese dialect speakers	12.38	11.41
Standard dialect speakers	14.39	14.24

Table 2 shows WER results for the same models but measured against test data categorized based on the speakers' dialect. Same as the WER results shown in Table 1, the WER value of the model with combined dialect lexicon is lower than that of the model with standard dialect lexicon. This shows that the model with the combined dialect lexicon has better performance.

4 Analysis and Discussion

We initially thought that the variety of pronunciations may create problems in data consistency, however, after doing some careful analysis, we found out that such pronunciation diversity can actually enrich our data. We decided to create a unique pronunciation symbol for each different pronunciation. Therefore, in the lexicon, we represent these different pronunciations by assigning varied phonemes composition to a word.

berbagai	b ax r b ah g ay
berbagai	bh ax r bh ah gh ay b
erbagai	bh ax r bh ah gh ey
berbentuk	b ax r b ax n t uh k
erbentuk	bh ax r bh ax n t uh k
berbentuk	bh ax r bh ax n t uh
berdetak	b ax r d ax t ah k
berdetak	bh ax r dh ax t ah k
berdetak	bh ax r dh ax t ah

Based on experimental results, the ASR model generated using this lexicon data was able to improve ASR system's accuracy. Before adding pronunciation variations to the lexicon, the ASR system could not recognize a number of words spoken by the Javanese dialect speakers. But after adding varied pronunciations in the lexicon, the ASR system was able to recognize those words.

5 Acknowledgement

The publication of this paper is part of a research project funded by the Productive Innovative Research Grant (RISPRO) LPDP Competition, Ministry of Finance of the Republic of Indonesia based on Letter of Agreement No. PRJ-31 / LPDP / 2020. This publication is also made possible due to the good cooperation provided by Solusi247 and the Harapan Kita Heart and Blood Vessel Hospital.

6 References

Adda-Decker, M., & Lamel, L. (2000). The use of lexica in automatic speech recognition. In Lexicon Development for Speech and Language Processing (pp. 235-266). Springer, Dordrecht.

Ajami, S. (2016). Use of speech-to-text technology for documentation by healthcare providers. The National medical journal of India, 29(3), 148.

Ali, A., & Renals, S. (2018, July). Word error rate estimation for speech recognition: e- WER. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 20-24).

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, A. M. (2019). Tata bahasa baku bahasa Indonesia.

Banay, G. L. (1948). An introduction to medical terminology I. Greek and Latin derivations. Bulletin of the Medical Library Association, 36(1), 1.

Chiu, C. C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., ... & Zhang, X. (2017). Speech recognition for medical conversations. arXiv preprint arXiv:1711.07274.

Goss, F. R., Blackley, S. V., Ortega, C. A., Kowalski, L. T., Landman, A. B., Lin, C. T., ... & Zhou, L. (2019). A clinician survey of using speech recognition for clinical documentation in the electronic health record. International journal of medical informatics, 130, 103938.

Gunarso, G., & Riza, H. (2016, December). An overview of bppt's indonesian language resources. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12) (pp. 73-77).

Gunarso, M., Uliniansyah, T., & Santosa, A. (2016, October). Development of a Speech Corpus for an Indonesian Text-to-Speech System. In 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA), Bali (pp. 26-28).

Gunawan, M., Nurfadhilah, E., Aini, L., Uliniansyah, M., Gunarso G., Santosa, A., & Junde, J. (2018). Uji Coba Korpus Data Wicara BPPT sebagai Data Latih Sistem Pengenalan Wicara Bahasa Indonesia. *Jurnal Linguistik Komputasional, 1*(2), 45-50.

Hoyt, R., & Yoshihashi, A. (2010). Lessons learned from implementation of voice recognition for documentation in the military electronic health record system. Perspectives in health information management/AHIMA, American Health Information Management Association, 7(Winter).

Kaldi: Data preparation. (n.d.). Kaldi. Retrieved February 25, 2021, from https://kaldi- asr.org/doc/data_prep.html

Kamajaya, I., Moeljadi, D., & Amalia, D. (2017, September). KBBI Daring: A revolution in the Indonesian lexicography. In Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference (pp. 513-530).

Laboratorium Kebinekaan Bahasa dan Sastra, Kementerian Pendidikan dan Kebudayaan. (n.d.). Daftar Bahasa Daerah di Indonesia. Laboratorium Kebinekaan Bahasa Dan Sastra. Retrieved April 25, 2021, from https://labbineka.kemdikbud.go.id/bahasa/daftarbahasa Popović, M., & Ney, H. (2007, June). Word error rates: Decomposition over POS classes and applications for error analysis. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 48-55).

Statistik, B. P. (2011). Kewarganegaraan, suku bangsa, agama, dan bahasa sehari-hari penduduk Indonesia: Hasil sensus penduduk 2010. Jakarta: BPS.

Uliniansyah, M. T., Nurfadhilah, E., Annisa, H., Gunawan, M., Aini, L. R., Santosa, A., ... & Riza, H. (2018, November). Utilizing Indonesian Allophones and Intraword Short Pauses Handling to Improve Performance of Indonesian Text-To-Speech. In 2018 International Conference on Asian Language Processing (IALP) (pp. 143-146). IEEE.

Uliniansyah, M. T., Riza, H., Santosa, A., Gunawan, M., & Nurfadhilah, E. (2017, November). Development of text and speech corpus for an indonesian speech-to-speech translation system. In 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA) (pp. 1-5). IEEE.

Uliniansyah, T., Riza, H., & Riandi, O. (2013, November). Developing corpus management system for Bahasa Indonesia the "Perisalah" project. In 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/ CASLRE) (pp. 1-4). IEEE.

Vogel, M., Kaisers, W., Wassmuth, R., & Mayatepek, E. (2015). Analysis of documentation speed using web-based medical speech recognition technology: randomized controlled trial. Journal of medical Internet research, 17(11), e247.

ANALYSIS OF LOCAL LANGUAGE DICTIONARY COMPLETENESS COMPONENTS: STUDY OF LEXICOGRAPHY

Nita Handayani Hasan, Sahril
Regional Office for Language in Maluku Province, Indonesia
nita.handayani@kemdikbud.go.id

Abstract

The practice of lexicography in Maluku Province was starting from the end of the 16th century by Georgius Everhardus Rumphius. Rumphius came to Ambon as a part of VOC. Then, he became interested in documenting the natural wealth in Ambon Island. His book, entitled Hebarium Amboinense, contains descriptions of plants, plants environment, and the function of plants. In addition to Rumphius, there was James T. Collins who intend to researching languages in Maluku Province. In 1977–1979 he collected vocabulary lists throughout Maluku Province (Collins, 2018). Besides the two experts, lexicography practices are also carried out by the owners of local languages. The phenomenon of missing local language raises the awareness to documentation regional languages. Unfortunately, dictionary which made by the owners of local languages has many lack because it is only based on personal knowledge. Two local language dictionaries that have been compiled by the community are Woirata dictionary and Banda Eli dictionary. This paper aims to describe the components in the local language dictionaries which compiled by the community. This paper used descriptive qualitative method. The data sources are the Woirata and Banda Eli bilingual dictionaries which compiled by the community. This research is a lexicographic study. The result of this study is the local language dictionaries which compiled by the community didn't include standard dictionary elements.

Keyword: Maluku, local language, dictionary

1. Introduction

Maluku Province is the province with the third largest number of regional languages in Indonesia. There are sixty-three regional languages (Bahasa, 2019). It is scattered all over the island. The large number of regional languages leaves its own problems. Many of it waiting to be documented and preserved.

The lack of attention from the regional government about local languages documentation has made regional languages increasingly marginalized. The local government has not yet considered about regional languages loss problem as a problem that must be resolved quickly. In fact, from a regulatory perspective, local government has an authority to maintain the existence of regional languages.

Local government carelessness of local languages existence doesn't kill the owners of local language spirit to document their local languages. The owners of local language takepreventive steps to maintain the existence of their local languages. One of the preventive steps which they take is making local language dictionaries.

Local language dictionaries construction in Maluku Province has been carried out since the 16th century by Georgius Everhardus Rumphius. Because of his admiration for the natural wealth in Maluku Province, Rumphius documented it in a list of plants names, environment, and uses. Besides Rumphius, there is a list of vocabulary and dictionaries of languages in Central Maluku which compiled by Sierevelt (1922), Niggemeyer (1952), Tauern (1928), dan Devin (1978) (Collins, 1986). Another lexicography product in Maluku was carried out by Petrus Drabbe (1932), a missionary from the Netherlands. He was writing the Fordata - Indonesian Dictionary (MSC, 2017). The high interest of foreign researchers and missionaries on documenting regional languages in Maluku Province shows that local language must be protected.

The awareness of regional language dictionary preparation is also owned by local language users. they are also worried about the extinction of their local languages. Because of that, they try to make dictionary. Two local language dictionaries compiled by the speaking community are the Woirata dictionary and the Banda dictionary.

Woirata language is the language spoken by people on Kisar Island, Southeast Maluku, Maluku Province. It is the only language in Kisar island which non-Austronesian style. The Woirata language is similar to the Fataluku language in Timor-Leste, and belongs to the East Timor subgroup of the Timor-Alor-Pantar (TAP) language family together with Makalero and Makasai (Schapper, 2012). Besides Woirata languages, people in Kisar island speak Meher, far southeast malay, dan Indonesian language. Young speakers of the Woirata language prefer to use Far Southeast Malay compared to Indonesian in their interactions (Engelenhoven, 2002). Woirata language is more often used for rituals, announcements, mythology, and interactions between local residents (Nazarudin, 2015). That's phenomenon made the traditional elders take the initiative to compile a Woirata language dictionary which can be used as the basis for teaching regional languages in schools on Kisar Island.

Banda language is the language spoken by the people in Banda Eli Village, Kei Besar Utara Timur District, and Elat Village, Kei Besar District, Southeast Maluku Regency, Maluku Province (Bahasa, 2019). Banda language is not spoken by the people on Banda Island. This is because in 1621, the native people of Banda Island were forced to leave their homes. The native people of Banda Island were victims of genocide, as a result of the trade war between Dutch and British trading companies (Miles, 2000). As a result of that incident, many indigenous people fled to the island of Kei Besar. They built new village in Eli and Elat village.

A view people which stay in Eli and Elat village, especially young generation, have gone abroad to get a better life. Because of that, young generation doesn't many time to practice Banda language. In order to maintain the existence of the regional language, the traditional elders compiled a simple Banda-Indonesian dictionary. Unfortunately, the dictionary is only circulating in certain circles.

Bilingual dictionary is the type of dictionary which uses to preservation regional languages. Bilingual dictionaries can help language learners in translation proses. It is a dictionary whose source language is not the same as the target language (Chaer, 2007). The source language is the language that will be explained in another language. The lexicographer in Indonesia prefer to make bilingual dictionary than monolingual and multilingual dictionary.

In preparing dictionary, careful planning is needed to get good dictionary. Good dictionary is a dictionary which easy to use and related to users. To make a dictionary easy to use, the dictionary must have criteria and components dictionary. Dictionary compilation components must also be available in local language dictionaries, so that local language learners can learning local language easily. Therefore, this paper will show a comparison of the completeness components in the Woirata-Indonesian dictionary and the Banda- Indonesian dictionary.

Dictionary, one of the lexicography products, considered as a reference in studying a language. The dictionary is a reference book which contains a list of words, arranged alphabetically, and contains descriptions of how to use the word.

No dictionary is perfect. This is because the contents of the dictionary will continue to change, even when a dictionary is finished printing. There is no perfect dictionary, but we can find a good one. A good dictionary tells the reader how to apply a word to speech, how to combine a word with a word, the types of text that are likely to appear, etc (Atkins, 2008).

A local language dictionary is usually a bilingual dictionary. It is because bilingual dictionary easily to use in language studying. Bilingual dictionary contains source language and target language. The source language is the language that is the input for the dictionary, or it is also known as the object of the inventory. Meanwhile, the target language is the language that describes the entry.

The completeness components dictionary are contained in the dictionary structure. Bergenholtz dan Trap (1995) divided five dictionary structures as macro, micro, frame, cross-reference, and access structures. Macro structure (macrostructure) is a list of word entries that must be arranged systematically and regularly. The macro structure of a dictionary is visible in the list of entries and subentries.

The microstructure is the information given to each word in the dictionary (Sterkenburg, 2003). Dictionary microstructure depending on the type of dictionary, and the information to be selected and used. Form of dictionary microstructure are entry / sublema, pronunciations, word classes, definitions or equivalents, synonyms, etc.

The frame structure consists of four main components, that is a table of contents, preface, introduction, and instructions for use (Bergenholtz, 1995). The table of contents is a component that shows the contents of the entire dictionary. The second component, preface contains information about the function of the dictionary, the subject of study, user groups, source and entry selection criteria, dictionary coverage, and other related information, including an introduction to the dictionary compiler. The third component, introduction contains background information on dictionary compilation decisions.

The fourth component of the frame structure is the dictionary user manual component. This component is divided into three categories, namely the type of information system; information organization and systematization; and linkage of information. Information type categories provide information on how to find each component of the dictionary, including finding entries, sublems, meaning of abbreviations, labeling. The second category, namely the organization and systemization of information, provides an overview of the techniques for presenting and arranging each piece of information. The third category, the linkage of information is related to the pattern of information relations in you which is indicated by a cross-referencing pattern. It can be concluded that the dictionary usage indicator component contains information on how to find each entry in the dictionary, entry structure, use of abbreviations, labeling, and how to cross-reference each entry.

The cross-referencing structure is divided into two, cross-referencing inside dictionary and outside the dictionary. The cross-referencing inside dictionary relates to the way the dictionary provides annotations for other items in the dictionary. The symbol used is the sign or the word *see*.

The last dictionary structure is the access structure. That structure makes dictionary easy to use. Access structure is a lexicographic indicator structure that shows the user on any information contained in the dictionary (Bergenholtz, 1995).

2. Method

The data sources in this study are bilingual Woirata-Indonesian dictionary (Kamanassa, 2018), dan bilingual Banda-Indonesian dictionary (B. Suatkap Djamudin; Wusorwut, n.d.). The dictionaries was made independently by speakers of Woirata and Banda languages.

The research data were obtained by reading and observing each dictionary structure. The dictionary structure is recorded by a data card and described according to the dictionary. The data analysis used was content analysis.

3. Result

The Woirata-Indonesia and Banda-Indonesian dictionary have different unique feature. This is because of the background that underlies the compilation of these dictionaries. Although both have the aim of preserving the local language, the Woirata- Indonesian dictionary has also been prepared as teaching materials of teaching regional languages.

The two dictionaries were compiled by local people, who had minimal knowledge of dictionary compilation. As a result, there are still many dictionary structures that have not been included in the two dictionaries. In addition, in terms of presentation techniques, the two dictionaries still have not applied the technique of presenting entries in accordance with the rules of dictionary compilation.

Even though there are still many shortcomings, the lexicography practice carried out by the people who own the Woirata and Banda languages must be well appreciated. They try to document their language as a treasure that must be preserved, so that later it can be passed on to the younger generation.

The results of the analysis of the components of the Woirata-Indonesian dictionary and the Banda-Indonesian dictionary are presented in table 1.

Table 1. Components of Dictionary

Dictionaries name	Macro Structure			Micro Structure					Frame structure				Cross	Acce
	list of words	word type	use	definition	synonym	Additional entry	use	equivalent	Table of contents	prologue	Instructions for use	introduction		
Woirata-Indonesia	√	-	√	√	√	-	√	√	√	√	-	√	-	√
Banda-Indonesia	√	√	√	√	-	√	√	-	-	-	-	-	-	-

4. Analysis and Discussion

Based on the analysis, it is known that the Woirata-Indonesian dictionary and the Banda-Indonesian dictionary do not have a complete structure. To get more information, the researcher will analysis dictionary completeness of the two dictionaries.

a. Macro structure

The Woirata-Indonesian dictionary and the Banda-Indonesian dictionary arrange entries based on the order of letters. The arrangement of entries in the two dictionaries is done alphabetically, from A to Z.

In the Woirata-Indonesian dictionary, the composition of the entries is arranged from the letter A to the letter Z, but there are several entries in a letter that are not arranged consistently. An example can be seen below.

kapa

kape

kapi

kamat

kamate

kamatana

Based on the example of the entry, it is known that there are inconsistencies in the ordering of letters in the entries. *Kamat, kamate*, and *kamatan*a entries are supposed to be arranged before *kapa* entries. Moreover, *kamatana* entry should be above the *kamatea* entry.

Banda-Indonesia dictionary more consistent in ordering entries than Woirata- Indonesia dictionary. Banda dictionary consistently lists entries in alphabetical order. However, there are some sublems that are not sorted consistently. The examples are,

mboo

mboo mbai

mboo fo

mboo mbee

mboo rau

mboo sala

mboo jaik

Sublemma *mboo* is not consistently arranged alphabetically. *Mboo fo* should be sorted after the *mboo mbee*, and *mboo jaik* should be sorted before *mboo rau*.

b. Micro structure

Woirata and Banda dictionaries have different microstructure completeness. Both contain word definition information, but not all dictionaries contain grammatical information, collocations, synonyms, and word usage.

Micro structure in Woirata dictionary more complete than Banda dictionary. The Woirata Dictionary presents the forms of calculation, homonyms, synonyms, compound words, absorption elements, names of days, names of months, structure of limbs, and cardinal directions. It is outside of the main sequence of entries.

The Woirata dictionary does not provide word classes, pronunciations, sublema forms, and examples of usage in the main entry. The sublema is not written attached to the main entry. Examples of writing entries in the Woirata dictionary:

Ta-wa'	(1) tambah; (2) semakin
Tawa huhutina	semakin bertambah
Tawa mahune	membesar-besarkan
Tawa tu'ure	bertambah berat
Tawa lapane	bertambah banyak
Tawanara	bertambah terang

Apart from the Woirata-Indonesian form, the Woirata dictionary also presents entries in the Indonesian-Woirata form. The entries in Indonesian-Woirata are also arranged in alphabetical order. Unfortunately, these entries are not equipped with word class information, pronunciation, and usage examples.

The microstructure in the Banda language dictionary is not as complete as the microstructure in the Woirata dictionary. The Banda language dictionary presents the form of entries, sublems, and definitions. Examples of writing entries in the Banda dictionary, namely

FUKAR	gunung
FUKAR TUTUNO	puncak gunung
FUKAR AU	gunung api
FUKAR WANDAN	kepulauan Banda, daratan Banda

The techniques for presenting entries in the Woirata and Banda dictionaries are very different than other dictionaries. The two dictionaries have not used the entry presentation technique that is usually used in other dictionaries.

c. Frame structure

The frame structure of the dictionary looks at the components that make up the dictionary. These components are the table of contents, introduction, introduction, and instructions for use. The table of contents is a section that contains all the information contained in the dictionary; the foreword contains information about the functions of the dictionary, the purpose for which the dictionary is compiled, the dictionary user group, and the scope of the dictionary; the introductory component contains the dictionary creation decisions.

The Woirata dictionary has a fairly complete frame structure. Although the Woirata dictionary does not have any instructions for using the dictionary, it does list the grammatical structure of the Woirata. This is inversely proportional to the Banda language dictionary. The Banda language dictionary has absolutely no frame structure.

If seen from the background, the purpose of preparing the Woirata-Indonesian dictionary is that in addition to documenting regional languages, the dictionary is also expected to be used as teaching material for teaching regional languages in schools. Therefore, many of the contents of the Woirata-Indonesian dictionary have been adapted to local language teaching in schools.

The Banda-Indonesian dictionary is compiled independently by Banda language speakers, so that the dictionary is only a collection of lists of Banda language entriestranslated into Indonesian. Physical form is very simple, and only circulated in limited circles.

d. Cross reference structure

Cross-references in a dictionary are related to the way a dictionary provides reference information to other entries in a dictionary. Cross-referencing is useful for knowing the recommended form, and the correct one in terms of writing or word form.

The Woirata-Indonesia and Banda-Indonesia dictionaries do not have cross- referenced structures. This is because the dictionary characters do not allow cross- referencing. The entries in both dictionaries are the correct form of entry in the writing. e. Access structure

The access structure is a lexicographic indicator that can be used by users to obtain information in the dictionary. Hausman and Wiegard (1989) in Setiawan (2009) reveals that there are two access structures, namely external and internal access. External access can be facilitated by providing information on the macro structure, in the form of alphabetic ordering of entries. Internal access can be obtained through meaning numbering which can assist in finding specific target entries.

The access structure in the Woirata-Indonesian and Banda-Indonesian dictionaries can be seen in the micro and macro structures. The dictionary which present many macro and micro structure can easier in use. The Woirata-Indonesian dictionary has a more complete macro and micro structure than the Banda-Indonesian dictionary. However, the Woirata-Indonesian and

Banda-Indonesian dictionaries are local language dictionaries whose have different purpose.

The Woirata-Indonesian dictionary has special sections to give additional information that can make it easier for dictionary users. These informations are part of the count, homonyms, synonyms, compound words, absorption elements, the names of the days, the names of the months, the structure of the limbs, the funds of the wind direction.

5. References

Atkins, B. T., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

B. Suatkap Djamudin; Wusorwut, F. (n.d.). *Kamus Bahasa Banda - Indonesia*. Bahasa, B. P. dan P. (2019). Bahasa dan Peta Bahasa di Indonesia. Retrieved from https://petabahasa.kemdikbud.go.id/provinsi.php?idp=Maluku

Bergenholtz, H., & Tarp, S. (1995). *Manual of Specialised Lexicography*. Amsterdam: John Benjamin Publishing.

Chaer, A. (2007). *Leksikologi dan Leksikografi Indonesia* (Pertama). Jakarta: Rineka Cipta.

Collins, J. T. (1986). The Historical Relationships of the Languages of Central Maluku, Indonesia. *Language*, *62*(2), 471. https://doi.org/10.2307/414707.

Engelenhoven, A. van. (2002). Verb Sequences in Melayu Tenggara Jauh: The Interface of Malay ang the Indigenous languages of Southwest Maluku. In *Between Worlds: Linguistic Paper in Memory of David John Prentice* (pp. 177--192). Pacific Linguistics. https://doi.org/10.15144/PL-529.cover.

Kamanassa, J. P., Wedilen, L., & Tamindael, O. (2018). *Kamus Bahasa Woirata*. Bandung: Alfabeta.

Miles, G. (2000). *Nathaniel's Nutmeg. How One Man's Courage Changed the Course of History*. London: Hodder and Stoughton.

MSC, P. D. (2017). *Kamus Bahasa Fordata*. Yogyakarta: Gunung Sopai.

Nazarudin. (2015). Causative constructions in Woirata , Kisar Island (Southwest Maluku , Indonesia). *Wacana*, *16*(1), 27–41. https://doi.org/10.17510/wjhi.v16i1.365

Schapper, A., Huber, J., & Engelenhoven, A. V.(2012). The historical relation of the Papuan languages of Timor and Kisar. *Language & Linguistics in Melanesia. Special Issue: On the History, Contact & Classification of Papuan Languages - Part I*, 59–87. Retrieved from http://www.langlxmelanesia.com/sanroque-loughnane387- 427.pdf%0Ahttp://www.langlxmelanesia.com/Special 4 robinson-holton59-87.pdf

Setiawan, T. (2009). Analisis Struktur Kamus Monolingual Bahasa Indonesia. *Litera*, *8*(2), 179–192.

Sterkenburg, P. V. (2003). *The History: Definition and History*. Amsterdam: John Benjamin Publishing.

THE LOCAL LANGUAGE DICTIONARY COMPOSITION STRATEGY IN PRESERVING ENDANGERED LANGUAGE

Nurul Qomariah
Regional Agency for Language in North Sulawesi Province, Indonesia
nurul.qomariah73@gmail.com

Abstract

According to the Basic Language and Literature Data page, National Agency for Language Development and Cultivation stated that the Ponosakan language in North Sulawesi Province is included in an endangered language. Refers to the data, it is very urgent to preserve of the Ponosakan language. One of the efforts in preserving for the endangered Ponosakan language requires concrete steps, namely documenting this language in a dictionary.

An Effort to preserve and maintain the existence of an endangered local language through the composition of a dictionary certainly require a special strategy. The arrangement needs to look at what things are needed, especially in the category of endangered languages, such as the Ponosakan language. The strategy for compiling the dictionary can be seen from various sides, both from the dictionary users as consumers and from the dictionary compilers as producers. Thus, a strategy for preparing local language dictionaries is absolutely necessary in order to produce a good and acceptable dictionary product. However, references that specifically discuss strategies that can be used to compile a dictionary of endangered languages are limited. In fact, a more detailed reference is needed to be used as a consideration in compiling a dictionary with the endangered language category in Indonesia. Therefore, a review of references on the compilation of a dictionary of endangered local languages is discussed in this paper to broaden the perspective in formulating a strategy for compiling a dictionary of endangered languages.

This study found fourteen strategies that can be used as a reference for compiling a dictionary of local languages that are endangered language. The strategy for compiling local language dictionaries is expected to be able to fill in gaps that have often occurred in local language dictionaries that previously existed. In addition, this strategy review is also expected to be a trigger for the compilers of the endangered language dictionary before the dictionary is launched into the community.

Keywords:

strategy, dictionaries composition, local languages, languages preservation, endangered languages.

I. Introduction

The study of this paper is an effort for researchers to explore the composition of local language dictionaries that are being concerned more about it. Especially with the Expert Group for Professional Service in Lexicography and Terminology (KKLP KI), which was initiated by the Head of National Agency for Language Development and Cultivation at the end of 2020 and is effectively applied to all Technical Implementation Units throughout Indonesia in 2021, including the Language Agency of North Sulawesi Province, where the author is a member of this KKLP KI. This opportunity is a trigger for further deepening in the field of lexicography, especially the compilation of local language dictionaries and of course the potential of local languages in the author's own working area, namely North Sulawesi Province.

The province of North Sulawesi has ten local languages spread across the speaker's territory. The ten languages are listed on the Language Agency's website, namely https://petabahasa.kemdikbud.go.id/

provinsi.php?idp=Sulawesi%20Utara. The details of the ten local languages in the North Sulawesi Province are (1) Bantik language; (2) Mongondownese; (3) Gorontalo language; (4) Malay; (5) Minahasan; (6) Tonsawang Minahasan; (7) Tonsea Minahasan; (8); Pasan language; (9) Ponosakan language; and (10) Sangihe Talaud language. In the past, Minahasa language in the North Sulawesi region was mapped by Riedel in 1858 in Manoppo (1983) there are eight languages, namely Tombulu, Tonsea, Tontemboan, Tolour, Tounsawang, Bantik, Bentenan or Pasan, and Ponosakan. Meanwhile, Adriani in 1925 was still in Manoppo (1983) divided two categories of Minahasa languages in the North Sulawesi region as native Minahasan, namely Tombulu, Tonsea, Toulour, Tountemboan, and Tounsawang. The non-native Minahasan, namely Bantik and Bentenan which are related to Sangir, Ponosakan—languages that are classified by Adriani as the old variety of Mongondow language, and Bajo language.

Of the many local languages that live in the North Sulawesi Province, one local language is registered on the Language Agency page, https://dapobas.kemdikbud.go.id/homecat.php?show=url/reglanguage&cat=5 which is categorized as endangered, namely the Ponosakan language. Referring to the online KBBI page—https://kbbi.kemdikbud.go.id/entri/ponosakan, it is stated that Ponosakan is the language spoken by the Ponosakan tribe, an ethnic group that inhabits the Minahasa Regency area. The Ponosakan language lives in the speech area of Southeast Minahasa Regency, precisely in Belang District. Ponosakan is an endangered language because there is only one village that is still using this language actively, namely Tababo Village, there are four fluent speakers between 70 and 92 years old in Lobel's findings (2016) during the 2015—2017 period.

The composition of the Ponosakan language dictionary so far as one of the language documentation efforts has not been carried out thoroughly. In fact, documentation of language through the composition of dictionaries is very necessary for languages with categories such as Ponosakan. Therefore, efforts are needed to compile local language dictionaries with the endangered category. However, the effort to compile a local language dictionary with the endangered category certainly requires a separate strategy—a careful plan to achieve specific targets (KBBI online). A special strategy is needed in the effort to compile a dictionary of endangered local languages, especially to preserve for local languages through language documentation. Therefore, this study seeks to examine strategies that need to be considered from several similar studies and references to the composition of local language dictionaries, especially those that are endangered, to be applied to the composition of local language dictionaries that have a similar category, namely endangered languages.

The local language dictionary's composition in general is followed by a flow that has been based on guidelines from upstream to downstream that have been compiled by experts, such as Sunaryo (2001) who specifically explained the composition of local language dictionaries and Setiawan (2015) who discussed the composition of dictionaries accompanied by his theory. These two theories complement each other, but the strategy for compiling a local language dictionary specifically for endangered languages is not mentioned in the Sunaryo and Setiawan guidelines. The strategy for compiling a dictionary of endangered local languages is very limited, so this research is important to do. Its application can be applied in preserving for the Ponosakan language which is categorized as endangered. Thus, this research will represent the strategies that need to be carried out in an effort to compose a dictionary of local languages that are endangered languages.

Based on a literature review on the compilation of existing language dictionaries, the formulation of specific strategies to develop endangered languages is still limited. In fact, there are a number of local languages in Indonesia which also have an endangered category in addition to the Ponosakan language, North Sulawesi, such as (1) Adang language, NTT; (2) the Kalabra language, West Papua; (3) Nedebang language, NTT; (4) Samasuru language, Maluku; (5) Bajau Tungkal Satu, Jambi; (6) Bayan language, Central Kalimantan; (7) Budong-Budong language, West Sulawesi; (8) Cia-Cia language, Southeast Sulawesi; (9) Emplawas language, Maluku; (10) Golic language, West Kalimantan; (11) Haruku language, Maluku; (12) Asilulu language, Maluku; (13) Ibu language, North Maluku; (14) Kayu Agung, South Sumatra; (15) Kayu Agung Lintang language, South Sumatra; (16) Kerinci language, Jambi; (17) Komodo language, NTT; (18) Kulawi, Central Sulawesi; (19) Lematang, South Sumatra; (20) Manyaan language, Central Kalimantan; (21) Devayan language (Simeuleuh), Aceh; (22) Besoan, Central Sulawesi; (23) Biak language, Papua; and (24) Mongondow (Bintauna), North Sulawesi. Therefore, this research is expected to offer a strategy for compiling local language dictionaries that are focused on endangered languages. Of course, this strategy is expected to be a way out for documenting 25 local languages in Indonesia with other endangered statuses from the 320 languages that have been registered on https://dapobas.kemdikbud.go.id/homecat.php?show=url/regbahasa&cat=5 (the National Agency for Language Development and Cultivation's linguistic data page).

The composition of the local language dictionary is described in detail by Sunaryo (2001:3-4) into several stages, namely preparation, data collection, data processing, data presentation, computerization, script preparation, editing, printing drafts, checking printouts, final editing, publishing, and distribution. The stages of Sunaryo's details in preparing local language dictionaries are very helpful for dictionary compilers to keep track of the completion of dictionary projects. However, the composition of dictionaries specifically for endangered languages is different from the dictionaries composition in general. This is not discussed by Sunaryo (2001) and Setiawan (2015) in detail, but Mosel (2004) discusses it in detail. Mosel stated that the main differences were mainly in the relatively short time for compiling the dictionary, the small number of drafting staff, and limited funding. This is as Jackson (2002: 161) stated that compiling a dictionary requires considerable investment, both in staff, equipment, discussion, and time. The dictionary project must run within budget and schedule. Therefore, a dictionary compilation project must be planned and managed properly. In addition, it is also necessary to involve people with various special knowledge and skills.

The composition of dictionaries, especially languages that are categorized as endangered, such as the Ponosakan language will require certain strategies that will be needed in the future for local languages in Indonesia with similar categories. Not only in terms of urgent time for language documentation efforts through the composition of a dictionary, which takes into account the limited active native speakers, but also considers the reliability of the dictionary compilation team who has knowledge and experience in documenting the language. In addition, it is necessary to use dictionary user research to accommodate future dictionary users (Lew and De Schryver, 2014). The strategy for compiling the dictionary can be seen from various sides, both from dictionary users as consumers and from dictionary compilers as producers. Thus, a strategy for compiling local language dictionaries is absolutely necessary in order to produce a good and acceptable dictionary product.

Ponosakan language dictionary has not been worked out optimally so far, both monolingual and bilingual dictionaries. Until this study was carried out, the composition of the Ponosakan language dictionary was still limited to reports on activities carried out by the dictionary compilation team of Language Center of North Sulawesi in 2014. However, unfortunately the composition of the Ponosakan language dictionary has not been continued at the publication stage so that it has not been exposed yet. There is only one Ponosakan language documentation activity that has been published in an international journal by Lobel (2016). Lobel compiled a phonological list of Ponosakan language entries in online dictionary form at http://talkingdictionary.swarthmore.edu/ponosakan/. The speaking dictionary is able to represent the pronunciation in Ponosakan language, although in a limited number. There is still a gap for the compilation of a dictionary of endangered languages, in this case Ponosakan language. It is still needed more entries to comprehend Ponosakan language in other format, not only speaking dictionary. This study offers a special strategy that can be used as a reference similar to the status of the Ponosakan language which is already endangered to be immediately documented through the composition of a dictionary. Therefore, this study offers a special strategy that can be used as a reference similar to the status of the Ponosakan language which is already endangered to be immediately documented through the compilation of a dictionary.

II. Classification of Compilation of Endangered Local Language Dictionary

There are five literature references that are reviewed and classified in this paper to formulate strategies in the composition of a dictionary of endangered local languages. The five references are Rehg and Cambell (2018), Mosel (2004), Ogilvie (2010), Kotorova (2016), and Ivanishcheva (2016). The experience of the lexicographers who specialize in compiling the endangered language dictionary is referred to by researchers to formulate strategies that can be applied in the composition of the endangered local language dictionary in Indonesia. It is hoped that the formulation of specific strategies for the composition of a dictionary of endangered local languages can be considered by the team that composed a dictionary in the similar languages category.

This section will classify some references of dictionary compilation of endangered languages that researcher has examined. Each reference will be parsed individually. The first reference, namely Rehg and Campbell (2018).

Some of the criteria for determining language endangered are described by Rehg and Campbell (2018: 1), namely; (1) absolute number of speakers—the fewer speakers, the less likely the language is to survive in the long term; (2) intergenerational transmission—if language is not learned by children in the traditional way, passed down from one generation to the next, it will essentially become endangered unless revitalization efforts prove successful, the greater the intergenerational transmission, the sadder the survival of the language will be; (3) a decrease in the number of speakers—the fewer the number of speakers, the more threatened the endangered of the language; (4) reduced use of domains—the less domains in which the language is used, the greater the chance of its endangered.

The loss of language which is the fate of language is increasing sharply and is happening everywhere. Rehg (2018:4) gives the example of California in 1850 having 100 American Indian languages, but only eighteen are still spoken today; nothing is learned by children in conventional ways. This is similar to the description of the Ponosakan language when the Ponosakan Language Dictionary Compilation Team from North Sulawesi Language Center documented the language in 2014. The number of speakers who were active in Ponosakan language at that time was estimated to be only around 150 people by aged 50 years and over, while those aged 50 years and under had no longer mastered the Ponosakan language. Ponosakan people aged 50 years and under use Manado Malay in their daily conversations. This is also not surprising because Manado Malay is the lingua franca in North Sulawesi. At that time, only a few people spoke Ponosakan fluently. The same thing was stated by Lobel (2016) when he documented the phonological of Ponosakan language two years after the team that compiled the dictionary of the North Sulawesi Provincial Language Center did the same thing.

Compiling a dictionary is the same as making a useful product, this is stated by Rehg (2018) who summarized five steps in creating a successful product. At least these five steps or stages need to be passed to compile a good dictionary, including a dictionary of endangered languages, namely (1) research; (2) initial planning; (3) design and construction; (4) distribution; and (5) support.

The first step in compiling a dictionary, according to Rehg (2018), is conducting research, both basic and applied research. Basic research includes linguistic knowledge, such as phonological orthography, phonology as a reference for the formation of pronunciation, morphosyntax by providing language labels, morphological knowledge to determine word formation, semantics as a reference for the formation of more than one definition, sociolinguistic knowledge to measure specifically the level of word use, the history of linguistics is necessary for the formation of etymology. However, the compilation of a dictionary of endangered languages is usually limited in scope and size. As for the scope of applied research, knowledge of the basics of lexicography, principles, and dictionary composition exercises is required.

The second step in compiling the dictionary formulated by Rehg (2018) is proper planning with considerations based on several things, such as dictionary users. Rehg (2018: 309) places the position of dictionary users as the main consideration in compiling a dictionary. The characteristics of dictionary users need to be known in order to prepare the right and appropriate dictionary for its users.

The type of dictionary to be compiled is also a determining part of planning considerations in compiling a dictionary. Rehg (2018) emphasizes that dictionaries of endangered languages are usually produced with a limited time frame and limited funding, sometimes even without funding. This consequence that makes the compilation of dictionaries with these conditions varied in scope—can be in the form of thematic dictionaries, focusing on one or several semantic domains, based on the corpus, or more comprehensive specialized dictionaries.

Another consideration that is included in planning the composition of the dictionary formulated by Rehg is the orthography that will be used. It should be noted about the pronunciation system used.

Another thing that needs to be considered in planning the compilation of the dictionary, namely the staff who will assist in the composition of the dictionary, both core staff and support staff. Rehg selected core staff who must come from the language community—native speakers—have fluency in the language. It is important to include elders/elders in the speaking community as part of the staff; they usually have a broad vocabulary and their outlook is richer than younger speakers. The support staff become consultants with specific areas of expertise, such as fishing, hunting, building, agriculture, medicine, etc. which are considered important in the community. The other support staff members may be involved as computer and recording experts. The team in the composition of the dictionary also requires a member who serves as the main editor—usually this is done by a linguist and the lead consultant. They are tasked with maintaining the consistency and final agreement contained in the course of a dictionary.

Other planning that needs to be taken into account in compiling a dictionary is choosing the right software as a database for a collection of dictionaries that have been compiled and adapted to the type of dictionary to be produced.

Another consideration that is included in the initial planning section in compiling the dictionary is the funding for the activity of compiling the dictionary. What kind of activity in the composition of this dictionary will be funded up to the stage of distributing the dictionary.

The third step of Rehg (2018) in compiling the dictionary, namely the construction and design of the dictionary. Rehg tended to direct the composition of bilingual dictionaries rather than monolingual dictionaries. The bilingual dictionary displays the target/target language which translates the source language—the local language into the target language. Consideration of the structure and design of the dictionary seen from the macrostructure, microstructure, and megastructure referred to by Rehg from Svensen (2009).

The fourth formulation of Rehg (2018) in the composition of the dictionary, namely distribution and support. It is also important to consider carefully how dictionaries are distributed and ultimately placed in the hands of their users. Rehg recommended that dictionaries of endangered languages compiler continued to support dictionaries that have been distributed to users. Continue to cooperate with the local education department for training in the composition of language dictionaries. Support for the resulting dictionary is needed for further dictionary work.

The second reference that discussed the composition of a dictionary of endangered languages is Mosel (2004). Mosel stated that the function of compiling an endangered language dictionary is, namely as a source of research and as a language repository for the speaking community. Therefore, the dictionary needed to be equipped with a thesaurus that covers various semantic fields such as kinship terms, names of animals and plants, terms related to the natural environment, material culture and social structure, activities, circumstances, and properties.

Some things that need to be considered in compiling a dictionary of endangered languages are conveyed by Mosel (2004) based on his experience, namely (1) the determination of the purpose of making a dictionary which must always be based on the identification of potential dictionary users and the purpose of using the dictionary. The point is, the dictionary must be able to fulfill the need and the interests of dictionary users; (2) selection of language variety. The determination of the variety of languages in the composition of the dictionary must be based on several careful considerations; (3) the time factor is as accurate as possible. The limited time that mostly of endangered languages dictionary compiler faced making Mosel suggested considering other alternatives to produce useful work that can be done in a short period of time. The alternative is by compiling a corpus-based dictionary and a thematic dictionary. Thematic dictionaries tend to be mini dictionaries or small dictionaries, but fulfill academic standards and can be useful to the language community and academics from various fields. Mini or thematic dictionaries compiled by Mosel, namely architecture and furniture of the Samoa language and followed by similar mini dictionaries on food, shipbuilding, fisheries, and important cultural practices in Samoa; (4) things related to orthography are almost similar to the choice of language variety. Determination of standard orthography is required for the compilation of an endangered languages dictionary so that they can be used in a standard manner; (5) grammatical information is contained as completely as possible to understand the abbreviations used in dictionary entries; (6) write a list of words as a provision for dictionary entries that can be used by the lexicon in searching for language data in the field. Mosel offered an active elicitation method by extracting word lists from linguistic data to find narrowly defined subject area words such as names of plants grown in the garden, types of houses, colors, etc. As for activities, it can be asked about the activities carried out when preparing food, such as ‘taking water’, ‘washing vegetables’, etc. Can also ask native speakers to look for basic words from certain semantic fields, such as whispering, shouting, asking, etc.; (7) The thematic approach is preferred in writing dictionary entries rather than the alphabetical approach. Thematic dictionaries are an important part of a dictionary project and don’t take too long. Thesaurus can be very useful for the development of teaching materials and other language maintenance measures. The efficiency of the thematic approach in compiling a dictionary of endangered local languages will be in the form of compiling a list of words that will be used as a provision for dictionary entries; (8) The writing of entries in the endangered local language dictionary can be adapted to the specificity of the language; (9) Team capacity building through internships and workshops.

The third reference on compiling a dictionary of endangered languages is Ogilvie (2010). In order to effort the language documentation through the compilation of dictionaries, including conservation and revitalization of endangered languages, Sarah Ogilvie suggested collaborative and innovative engagements between members of the academic community, members of the language community, indigenous communities, and non-governmental organizations. This effort is important to preserve cultural diversity and knowledge systems that can be encoded in dictionaries, not just a by-product of describing the grammar of a language.

In compiling the dictionary, Ogilvie (2010) divided it into three categories based on the level of language danger around the world, namely dictionaries for language preservation, dictionaries for language maintenance, and dictionaries for language revitalization. Following are the details of each category of the dictionary; (1) the dictionary for language preservation is exemplified by the Aslian language in the equatorial forest of Malaysia. The lexigographer of Niclas Burenhult and team focuses on the unique ethnobiological knowledge of forests and how to create sustainable livelihoods from them. Burenhult faced a complex decision regarding the order of entries and chose not to sort the main words alphabetically, but according to the manner and place of articulation in left-to-right rather than rhyming order; (2) a dictionary for language maintenance is exemplified by Ogilvie with a successful dictionary program called Free Electronic Dictionary. A dictionary project of the endangered Australian Aboriginal languages by James Mc Elvenny and Aidan Wilson at the University of Sydney. A small dictionary easily accessible via mobile by Aboriginal community people.

Small thematic dictionaries in the field of semantics have been initiated in the lexicography of endangered languages. This mini dictionary is perfect for breaking down comprehensive dictionary tasks. These dictionaries can give language community speakers instant access to their language dictionaries for use in schools and the general public. The mini dictionaries are a collaborative effort between older speakers assisting with editing and younger speakers checking for clarity of entries, as well as children providing feedback on the lexical scope of the dictionary (for example, Teop children in Papua collect shells they find missing in the dictionary, first draft of the book—shell dictionary).

Ulrike Mosel and Ruth Spriggs who compiled Teop mini-dictionary found that collaborative lexicographic activities can promote language awareness and pride of young speakers and demographics targeted for language maintenance or revitalization are successful. Involving speech communities in the composition of mini dictionaries is the result of lexicographical work with tangible results rather than waiting years for a comprehensive dictionary to be completed. This lexicographical work also showed the commitment of the lexicographer and team to the maintenance and revitalization of language in society; (3) a dictionary for language revitalization is prepared to answer quite complex issues related to endangered, literacy rates, and opportunities for capacity building and empowerment of community members to revitalize their language. Added pressure was also faced by the dictionary drafting team to finish quickly before the last speaker died. Dictionary of Christine Beier and lev Michael who compiled the Iquito Dictionary in the Amazon, Northern Peru advocated a team-based and community participatory approach to assist in rapid data collection.

The fourth reference to the compilation of endangered languages is the experience of Kotorova (2016). Elizaveta Kotorova stated that the creation of dictionaries of minority languages often goes beyond purely lexicographical work, becoming a theoretical and practical scientific endeavor and the main means of preserving indigenous languages and cultures. Kotorova described his experience when compiling a dictionary for a minority language—Ket language of Central Siberia which has its own peculiarities.

The first task that Kotorova did before starting to create a dictionary was to identify the target users by deciding who would be the potential users of the dictionary to be compiled. Kotorova stated that the decision to identify the target user will play an important role in choosing what should and should not be included in the dictionary. Kotorova added that when compiling a minority language dictionary, there were three possible choices regarding potential users, namely a dictionary for speech communities, a dictionary for academics, or a dictionary to accommodate both.

The second task that is no less important in compiling a dictionary according to Kotorova (2016) is collecting language material and compiling basic vocabulary. Kotorova formulated the opinion of linguists in making the initial word list as follows; (1) translate a list of the most frequent words from a common language, in this case English. Although this method is simple and easy enough, but this word list will not represent the source language lexicon and will lose the culture-specific concept because it has no translation equivalent in the source language; (2) extracting word lists from the corpus, but this is only applicable if the source language has corpus data, and usually does not exist in the source language, an endangered language; (3) thematic gathering of native speakers to find narrowly defined subject area words, such as color terms, layout, etc. This method helped to reveal the basic and culture-specific words of the source language.

The third task described by Kotorova (2016) in compiling a dictionary is making dictionary entries. The composition of each dictionary entry is determined by the potential purpose of the dictionary user. The semantic information included in the compilation of dictionaries and other types of information allows the introduction of a number of necessary parameters. Korokova emphasized that each dictionary entry includes two very important elements, namely the citation form and the commentary.

The fifth reference for compiling a dictionary of endangered languages is the experience of Ivanishcheva (2016). Olga N. Ivanishcheva stated that the lexic dictionary of endangered languages is a unique lexicographic product. Many factors have to be taken into account when compiling a dictionary like this, such as the purpose of the dictionary, grammatical features, choice of spelling if there is no standard spelling. However, there are very important things that need to be considered when compiling a dictionary of endangered languages, namely the type of dictionary to be created and the needs of the dictionary user. According to Ivanishcheva, these two things are the main factors that guide the selection of information to be used in the entry. It is also important to consider the level of language used in the definition of dictionary entries. The choice of language used in the definition of a dictionary entry must be match to the language level of the dictionary user.

All kinds of information are demanded by foreign language dictionary users, both spelling and pronunciation of individual words, collocations and pronunciations. A dictionary intended for speakers of a language and for those learning it also show some differences. In the latter case, there is a greater need for information on the grammatical characteristics of the word, its connotations and uses, including stylistic features. In either case, it is important to consider that the language level used in the definition of any word in the dictionary entry must be match the user's language level, or they will not understand it.

User type is even more important for dictionaries as they are for students. Learner dictionaries are usually divided into two groups based on their recipients: 1) dictionaries for foreigners (ie those whose languages are foreign/non-native); 2) dictionaries for native speakers who study their mother tongue as a subject (children school and students).

User type is even more important for dictionaries as they are for students. Learner dictionaries are usually divided into two groups based on their recipients: 1) dictionaries for foreigners (ie those whose languages are foreign/non-native); 2) dictionaries for native speakers who study their mother tongue as a subject (schoolchildren and students). Ivanishcheva states that B. Svensén includes these types of dictionaries in various oppositions: dictionaries for general use versus dictionaries for study; dictionary for adults – dictionary for children. A special feature of dictionaries for study purposes is that users of those dictionaries have limited possibilities for using the contents of the dictionary. The purpose of studying dictionaries is as an effective communication tool. A dictionary for children should not be just a shortened version of a dictionary for adults (Svensén, 1987 p. 20).

III. Formulation Strategy in Composition a Dictionary of Endangered Languages

Based on the results of the literature review on the five experiences of the compilers of the endangered language dictionary, namely Kenneth L. Rehg, Ulrike Mosel, Sarah Ogilvie, Elizaveta Kotorova, and Olga N. Ivanishcheva, the researcher formulated or offered several strategies that can be applied to the compiling of language dictionaries by endangered category. The formulation of the strategy in the composition of this endangered language dictionary will be presented in sequence as the process of compiling a dictionary. The following is the formulation of the strategy in the composition of the endangered language dictionary.

3. 1 Identify the target users of the dictionary

The strategy in compiling a dictionary of endangered languages begins with identifying the target users of the dictionary. This was done at the outset by Kotorova before compiling the dictionary. The decision to identify who the dictionary users are will have implications for the dictionary entries to be compiled. Ivanischeva also confirmed the same thing, namely that dictionary users must be prioritized in the composition of dictionaries so that the identification of dictionary users will guide dictionary authors in choosing the information to be presented in dictionary entries. The identification of potential dictionary users was also put by Mosel at the beginning of the compilation of the dictionary together with the determination of the purpose of making the dictionary. Likewise with Rehg, he placed the position of dictionary users as the main consideration in compiling a dictionary.

Dictionary users should be the main consideration in making a dictionary, this is stated by Kwary (2018). Errors in identifying the user can turn the dictionary into just a display without much benefit. Therefore, the target user of the dictionary is the first thing to get into the strategy for compiling a dictionary of endangered languages.

3. 2 Set the purpose of creating a dictionary

The purpose of making a dictionary must be always based on the identification of potential users of the dictionary and the purpose of using the dictionary. This is stated by Mosel and Kotorova. In essence, the needs and interests of dictionary users must be fulfill by the dictionary that will be compiled. Based on the experiences of Rehg, Mosel, Ogilvie, and Kotorova, the compilation of a endangered languages dictionary tends to be thematic, focusing on one or more semantic domains, based on a corpus, or a more comprehensive specialized dictionary.

3. 3 Set the schedule for compiling the dictionary as accurately as possible

The compilation of a dictionary is like a train carriage that runs slowly but surely on its tracks, starting from one station to the next. Scheduling of dictionary composition activities must be arranged in a planned manner because most endangered language dictionary compilers have limited time. Rehg and Mosel suggested setting the schedule as best as possible. Mosel suggested considering other alternatives to produce useful work that can be done in a short period of time. The alternative is by compiling a corpus-based dictionary and a thematic dictionary. Thematic dictionaries tend to be mini dictionaries or small dictionaries, but fulfill academic standards and can be useful to the language community and academics from various fields.

3. 4 Organize funding for dictionary preparation activities properly

Usually dictionaries of endangered languages are produced with limited funding, Rehg and Mosel pointed out. Therefore, both of them suggested to be able to arrange the funding of dictionary preparation activities properly, including how this dictionary preparation activity will be funded up to the dictionaries distribution stage.

3. 5 Select a staff or dictionary compilation team by involving the language community

Rehg selected core staff who must come from the language community—native speakers—have fluency in the language. It is important to include elders/elders in the speaking community as part of the staff; they usually have a broad vocabulary and their outlook is richer than younger speakers. The support staff become consultants with specific areas of expertise, such as fishing, hunting, building, agriculture, medicine, etc. which are considered important in the community. Other support staff members may be involved as computer and recording experts. Team members on the dictionaries are also required to serve as lead editors—usually this is done by a linguist and lead consultant.

The involvement of local language communities to assist in the composition of their local language dictionaries is also a separate strategy in compiling of endangered local language dictionaries. As Mosel did by providing training to local residents to help the team work. If possible and successful, local people will be able to continue to develop their own dictionaries.

The involvement of the local language community is limited by Mosel (2006:81), namely researchers and local language workers must always have a clear idea of what kind of work needs to be done and when it needs to be done. Therefore, the team jointly organizes their work based on the following points; (1) identify the types of activities required to produce documentation works; (2) discuss who will do what; (3) make a work plan by placing various activities into a certain order and allocating a certain time for each; (4) try to stick to the work plan; get one thing done before doing the next; (5) evaluate the work plan and revise it.

3. 6 Determine standard ortography correctly

Determination of standard orthography is needed for the composition of an endangered languages dictionary so that they can be used in a standard way. Moses emphasized this. This orthography needs to be agreed upon as a reference for the formation of pronunciation. Rehg emphasized the need to pay attention to the pronunciation system that will be used in compiling the dictionary.

3. 7 Provide complete grammatical information

Rehg and Mosel stated the grammatical information contained as complete as possible to understand the abbreviations used in the dictionary entries.

3. 8 Write a word list guide

Mosel stated that dictionary compilers need to write a list of words as a provision for dictionary entries that can be used to find language data in the field. Mosel offers an active elicitation method by extracting word lists from linguistic data to find narrowly defined subject area words, such as names of plants grown in the garden, types of houses, color variations, etc. As for activities, it can be asked about the activities carried out when preparing food, such as 'taking water', 'washing vegetables', etc. Can also ask native speakers to look for basic words from certain semantic fields, such as whispering, shouting, asking, etc.

3. 9 Pay attention to the structure and design of the compilation of the dictionary

Structure and design in the compilation of the dictionary need to be considered by looking at the macrostructure, microstructure, and megastructure referred to by Rehg from Svensen (2009). Macrostructure refers to information in an alphabetically arranged dictionary represented by a head word or an entry. The microstructure refers to the internal composition of dictionary entries, such as head words, alternative spellings, pronunciation, usage labels, definitions, phrases or example sentences, etymology, cross-referencing, and semantic domains.

The megastructure of a dictionary refers to all the components of the dictionary, including the front and back of the dictionary. The front of the dictionary includes a cover page, introduction, table of contents, names of contributors, dictionary usage information, a list of abbreviations or symbols, and a language map. Other important things that need to be on the front of the dictionary, namely pronunciation guide information, instructions for using the dictionary, information about spelling conventions, including word formation information, explanations of labels used, and levels of usage.

The back of the dictionary includes some information, such as a thesaurus containing a list of words that have specific semantic areas, grammar, ethnography, illustrations accompanied by descriptions (such as parts of traditional houses, traditional boats, etc.), maps, and names of regions. In essence, the back displays information about language and culture.

3. 10 Choose the right software as database

Rehg stated that the right software should be chosen as a database of a collection of dictionaries that have been compiled and adapted to the type of dictionary to be produced.

3. 11 Adapt the writing of dictionary entries to the specifics of the source language

The writing of entries in the dictionary of endangered local languages can be adapted to the peculiarities of the language.

3. 12 Conduct team knowledge capacity building

Capacity building for the team's knowledge of dictionary compilation techniques can be done through internships and workshops. Mosel suggested this is done by the compilers of the dictionary.

3. 13 Consider distribution of dictionaries to users

It is also important to consider carefully how dictionaries are distributed and ultimately placed in the hands of their users. Rehg recommended that dictionaries of endangered languages continue to support dictionaries that have been distributed to users.

3. 14 Cooperate with the local community after the dictionary project is over

Rehg menyarankan agar tetap menjalin kerja sama dengan masyarakat meskipun proyek penysunan kamus telah usai. Tim penyusun kamus dapat bekerja sama dengan departemen pendidikan setempat untuk pelatihan penyusunan kamus bahasa. Dukungan terhadap kamus yang dihasilkan tetap diperlukan untuk keberlanjutan pengerjaan kamus selanjutnya.

Rehg suggested that it should be continued to cooperate with the community even though the dictionary compilation project was over. The dictionary development team can work closely with the local education department for training in the compilation of language dictionaries. Support for the resulting dictionary is still needed for the continuation of further dictionary work.

IV. Conclusion

The strategy for compiling endangered local language dictionaries is still limited. In fact, local languages with an endangered status such as Ponosakan language in North Sulawesi need to be immediately documented through a dictionary. The strategy for compiling a qualified dictionary emerges from the lexicographers who have carried out dictionary compilation activities in the field with various obstacles encountered in the field. These various good practices carried out by lexicographers who concentrate on the field of endangered languages have led to strategies that can be adapted to the subsequent efforts to compose a dictionary of endangered local languages. The theoretical basis for the compiling dictionaries practiced has been introduced by experts, but the compilation of endangered local languages dictionaries requires a separate strategy based on the good practices of lexicographers.

These strategies were born from field experience supported by the basic theory of dictionary composition. Unfortunately, the strategy for compiling an endangered local language dictionary has not been mapped properly, so some efforts are needed to map it. Mapping the strategy for compiling an endangered local language dictionary will certainly help the team of developing similar dictionaries in the future.

This study examines various references from the good practice of lexicographers who concentrate on compiling dictionaries of endangered languages. Based on the results of this study, this study resulted in the formulation of strategies in the composition of the endangered language dictionary as follows; (1) identify the target users of the dictionary; (2) set the purpose of creating a dictionary; (3) set the schedule for compiling the dictionary as accurately as possible; (4) organize funding for dictionary preparation activities properly; (5) select a staff or dictionary compilation team by involving the language community; (6) determine standard ortography correctly; (7) provide complete grammatical information; (8) write a word list guide; (9) pay attention to the structure and design of the compilation of the dictionary; (10) choose the right software as database; (11) adapt the writing of dictionary entries to the specifics of the source language; (12) conduct team knowledge capacity building; (13) consider distribution of dictionaries to users; (14) cooperate with the local community after the dictionary project is over.

V. Reference

Atkins, B.T. & Rundell, Michael. 2008. *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Ivanishcheva, Olga N. 2016. Dictionaries of Critically Endangered Languages: Focus on Users. Dalam *Journal of Linguistics*, 2016, Vol. 67, Nomor 1, Halaman 73 – 86. https://sciendo.com/pl/article/10.1515/jazcas-2016-0012

Jackson, Howard. 2002. *Lexicography: An Introduction.* London: Taylor & Francis Routledge.

Kotorova, Elizaveta. 2016. Dictionary for A Minority Language: The Case of Ket. Dalam *Euralex*. 2016. Halaman 129—137. https://euralex.org/publications/dictionary-for-a-minority-language-the-case-of-ket/ dan https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_010_p129.pdf

Kwary, Deny Arnos. 2018. The variables for drawing up the profile of dictionary users. *Lexicography: Journal of ASIALEX*, *4*(2), 105–118. https://doi.org/10.1007/s40607-017-0030-x

Laksana, I Ketut Darma. 2014. *Manual Leksikografi: Metode dan Teknik Penyusunan Kamus.* Denpasar: Udayana University Press.

Lobel, Jason William. 2016. *Notes from the Field: Ponosakan: The Sounds of a Silently Dying Language of Indonesia, with Supporting Audio.* https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/24700/lobel_noaudio.pdf

Manoppo-Watupongoh, Geraldine Yvonne Jane, 1983. *Bahasa Melayu Surat Kabar di Minahasa pada Abad ke-19.* Disertasi. Fakultas Ilmu Budaya, Universitas Indonesia.

Mosel, Ulrike. 2004. Dictionary Making in Endangered Language Communities. Dalam Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 39—54. London: SOAS. School of Oriental and African Studies. https://www.mpi.nl/lrec/2002/papers/lrec-pap-07-ictionary_Endangered_SpComm.pdf

Mosel, Ulrike. 2006. Fieldwork and Community Language Work. Dalam *Essentials of Language Documentation.* Disunting oleh Jost Gippert, Nikolaus P. Himmelmann, dan Ulrike Mosel. Halaman 67—85. Berlin: Walter de Gruyter.

Ogilvie, Sarah. 2010. Lexicography and Endangered Languages: What Can Europe Learn from the Rest of the World? Dalam *Euralex*, 27—46. https://www.euralex.org/elx_proceedings/Euralex2010/000_Euralex_2010_01_Plenary_OGILVIE_Lexicography%20and%20Endangered%20Languages_What%20Can%20Europe%20Learn%20from%20the%20Rest%20of%20the%20Wor.pdf

Rehg, Kenneth L and Campbell, Lyle. 2018. *The Oxford Handbook of Endangered Languages*. New York: Oxford University Press.

Robert Lew, Gilles-Maurice de Schryver. 2014. Dictionary Users in the Digital Revolution. Dalam *International Journal of Lexicography*, Volume 27, Issue 4, December 2014, Pages 341–359, https://doi.org/10.1093/ijl/ecu011

Setiawan, Teguh. 2015. *Leksikografi.* Yogyakarta: Penerbit Ombak.

Sunaryo, Adi. 2001. *Pedoman Penyusunan Kamus Bahasa Daerah.* Jakarta: Pusat Bahasa.

Svensen, Bo. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary- Making.* New York: Cambridge University Press.

Online:

https://kbbi.kemdikbud.go.id/entri/ponosakan

Thank-you note

- My sincere thanks go to several assessors in the previous abstract.
- Thank you to the committee for being patient in providing additional time for me to complete the writing of this paper.
- Thanks to Ibu Dora Amalia, Pak Azhari Dasman Darnis, Mbak Dewi Puspita, Mbak Nurul Masfufah, and friends at KKLP of Lexicography and Terminology, National Agency for Language Development and Cultivation who have given confidence and enthusiasm in pursuing lexicography activities.
- Thanks to the compilers of endangered languages dictionaries around the world. Keep up the work!

SANTI-MORF DICTIONARIES

Prihantoro
Lancaster University, United Kingdom & Universitas Diponegoro, Indonesia
prihantoro@lancaster.ac.uk

Abstract

This paper highlights the role of dictionaries used in SANTI-morf (*Sistem Analisis Teks Indonesia – morfologi*), a multi-module pipeline that performs annotation for Indonesian corpus at morpheme level, built using Nooj (Silberztein 2003; 2016). SANTI-morf dictionaries, together with other SANTI-morf components, enable the system to tokenize each word in the target corpus into morphemes (e.g., cliticised and non-cliticised roots, affixes, reduplications) and associate these morphemes to their corresponding tags. Each entry in the SANTI-morf dictionary is encoded with a tag composed of Morphological Analysis (MA) labels and in most cases are combined with System Implementation (SI) labels. MA labels consist of formal and functional morphological criteria labels and can be used for searching the annotated corpus (e.g., root's Part of Speech (POS) labels. SI labels are used for system implementation and are mostly the interests of developers rather than end-users. They include morphotactic and morphophonemic constraint labels, which are processed when the monomorphemic entries in dictionaries work together with SANTI-morf grammars (rules).

Keywords: SANTI-morf, corpus, dictionary, grammar

1. SANTI-morf

Malay is genetically affiliated with the Austronesian language family. Over time, it has developed into different language varieties throughout Southeast Asia. Some of these varieties are standardised and serve as the official languages of a number of countries in this region (Indonesia, Malaysia, Brunei, Singapore).

Indonesian is one of the standardised Malay varieties, used as the national as well as the official language of the Republic of Indonesia. Lewis (2009) notes that Indonesian is spoken by almost 200 million speakers[1]. This constitutes Indonesian as the most widely used standardised Malay variety among other varieties spoken in Southeast Asia Polymorphemic words in Indonesian can be formed using a variety of morphological processes such as affixation, compounding, reduplication, cliticisation, or the combination thereof. Such processes can be analysed using computational morphology tools specifically designed for Indonesian morphology.

Pisceldo et al. (2008) create a Two-Level Morphological Analyser for Indonesian. Later, Larasati et al. (2011) build MorphInd, presented as an advancement of Pisceldo et al.'s tool. I review MorphInd's morphological annotation scheme and suggested some improvements (see more details in Prihantoro (2021:in press). To implement these suggestions, I create SANTI- morf, a new morphological analysis system for Indonesian text. SANTI-morf achieves 99% precision and recall on a testbed. The discussion of SANTI-morf is fully presented in my PhD thesis (Prihantoro 2021: forthcoming). The system itself is already available for use[2].

[1] Per 2010 national census, the population in Indonesia was at 230+ million. As of 2020 national census (the most recent), the population is at 270+ million. Thus, the number of Indonesian speakers are likely to improve.

[2] http://www.nooj-association.org/resources.html

SANTI-morf[3] is a rule-based text analyser for Indonesian which fully tokenises and annotates Indonesian words at morpheme level, not word level. SANTI-morf adopts a morphological annotation scheme devised by Prihantoro (2019). Dictionaries and grammars are two core components of SANTI-morf. These resources are grouped into four modules: the Annotator, the Guesser, the Improver, and the Disambiguator (see Prihantoro 2021: forthcoming). SANTI-morf is implemented using NooJ[4] (Silberztein 2003; Silberztein 2016), a finite-state based text analyser program.

Once a text is annotated using SANTI-morf, a user can search a morpheme (or a combination of morphemes), based on several criteria: the morpheme(s) linguistic unit, formal and functional morphological categories, or their combinations.

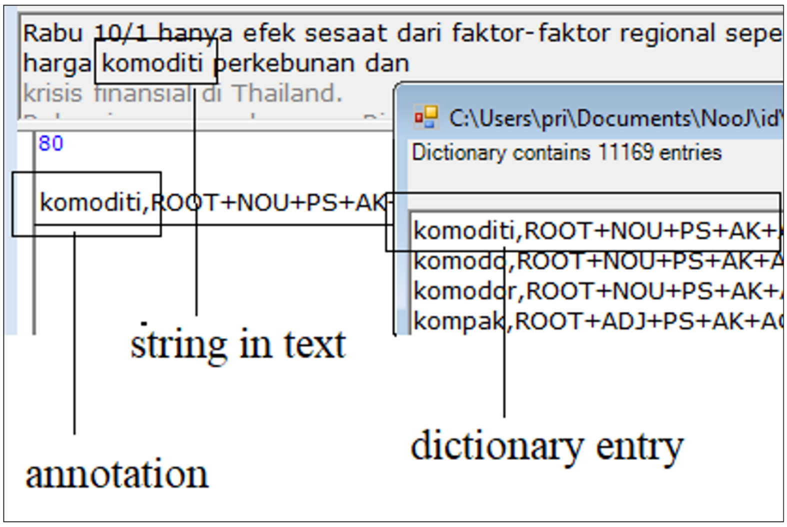
SANTI-morf may contribute to applications in different fields such as informatics, corpus linguistics, or lexicography. The application of SANTI-morf to support lexicographic works is demonstrated in section 4 of this paper. There is a wide range of aspects of SANTI- morf to discuss, but in this paper, I will focus[5] on describing the architecture of SANTI-morf dictionaries.

2. Dictionary entry

Dictionary entry is an important lexicographic component, using which, human users (e.g., language learners, linguists) can retrieve further linguistic information. Whether the dictionary is printed or electronic, dictionary entries are of great significance that helps human users interact with the dictionary.

Certain dictionaries do not directly target human users. Instead, they serve as resources for Natural Language Processing (NLP) applications, such as automatic text analysers, question answering system, predictive model, and many others. SANTI-morf dictionaries fall into this category.

When SANTI-morf system detects a string of characters in a text, it will always first perform a cross-examination with SANTI-morf dictionary entries before checking other types of resources (i.e., grammars). When matches are found in the dictionaries, SANTI-morf will annotate the string based on the labels encoded in the corresponding dictionary entries.



Rabu 10/1 hanya efek sesaat dari faktor-faktor regional sepe
harga komoditi perkebunan dan
krisis finansial di Thailand.

80

komoditi,ROOT+NOU+PS+AK

C:\Users\pri\Documents\NooJ\id\
Dictionary contains 11169 entries

komoditi,ROOT+NOU+PS+AK+
komodo,ROOT+NOU+PS+AK+A
komodor,ROOT+NOU+PS+AK+
kompak,ROOT+ADJ+PS+AK+A

Figure 1. An annotation based on a match found in one of the SANTI-morf dictionaries

[3] SANTI-morf is designed to be one of the core components of SANTI, a multi-module NLP pipeline to analyse Indonesian texts at various levels (morphology, morphosyntax/POS tagging, syntax/parsing, semantics, pragmatics, discourse, etc.).

[4] http://www.nooj-association.org/downloads.html

[5] SANTI-morf overall system (dictionary, grammar, module, configuration) is discussed in more details in Prihantoro (2021: forthcoming).

A dictionary file in SANTI-morf can be described as a file containing a collection of entry lines. Each entry line contains an entry and the corresponding tag (one or more labels), delineated by a comma. At this point, let's solely focus only on the entry. In all the examples in this section, I replace all the tags with an arbitrary code TAG. For example, the entry line below includes an entry *ikan* 'fish', whose actual tag is replaced by TAG.

(1) ikan,TAG

In terms of the number of morphemes, the entries can be divided into two categories: monomorphemic and polymorphemic. For instance, *getar* '(generic) vibrate' is a monomorphemic entry, but *gemetar* '(body part) tremble' is polymorphemic. Note that polymorphemic entries are reserved for words that are created using non-productive morpheme such as infix -*em*-[6]. Words produced by productive morphemes such as -*an* in *getaran* 'vibration', or *ber*- are analysed using the combination of dictionaries and grammars (or rules). SANTI-morf dictionaries and grammars are the core components of SANTI-morf, but in this paper, we will only focus on discussing the architecture of SANTI-morf dictionaries.

(2) getar,TAG

(3) gemetar,TAG

In terms of orthography, an entry may fully consist of letters, as most morphemes normally occur in texts. However, an entry can also be a non-letter symbol. This might seem trivial from a linguistic standpoint, but these characters are present in many actual texts. For example, chemical compounds are often written in codes that combine letters and numbers (e.g., h2so4) rather than what it is commonly referred to (e.g., sulfuric acid).

(4) h2so4,TAG

Emoticon (or emoji) may also serve to illustrate the use of non-letter symbols in actual texts, such as the combination of a colon and a closing bracket :) for a smiley emoticon. We see that the colon in the smiley emoticon entry line below is preceded by a backslash. This is because some non-letter characters are not allowed to be used alone as an entry in Nooj, such as colon character (:). To enforce this character as an entry, we must precede this entry with a backslash. Thus, instead of (:) , a colon entry is written as (\:). This differs from a question mark, which can be used independently without having to be escaped using backslash.

(5) \:),TAG

(6) ?,TAG

Some non-letter characters have special meaning. For example, the equal sign in the entry *kura=kura* allows SANTI-morf to recognise both *kura-kura* 'turtle' and *kura kura* 'turtle'. This is a useful feature as in the running texts, the hyphen is often dropped and is replaced with space. The presence of hyphen and space in the above examples give the impression that these entries are polymorphemic, whereas *kura-kura* and *kura kura* are actually monomorphemic. By fully encoding *kura=kura* as one entry line in SANTI-morf dictionary, SANTI-morf will never analyse *kura-kura* or *kura kura* as two tokens.

[6] Note that infixes are considered unproductive morphemes in Indonesian

(7) kura=kura,TAG

Another aspect of the dictionary entry is case sensitivity. If an entry is written in full lowercase, it is very flexible. The full-lowercase entry such as *bagian*, can be used to annotate all instances of *bagian*, namely *bagian, Bagian*, or *BAGIAN*. However, if an entry is written in uppercase, or include uppercase character(s), the matching will be very strict. For instance, the entry *Bandung* 'name of a city in Indonesia: always begins with an uppercase', will always strictly match Bandung, which begins with an uppercase in the text, with but not *bandung* which is written in full lowercase.

(8) bagian,TAG

(9) Bandung,TAG

3. Dictionary tag

As it has been discussed previously, an entry line in SANTI-morf dictionary is composed of an entry and a tag. In the previous section, I replaced the tag with an arbitrary label, TAG as we were focusing on discussing the entry. In this section, we will discuss the format of SANTI- morf tag in more detail. In SANTI-morf, a tag is defined as a label or a sequence of labels connected using a plus symbol (+). This can be illustrated by the tag for a dictionary entry *pohon* 'tree' below. The tag is composed of 8 labels overall; the plus (+) symbol is used to connect one label to another. The labels accompanying each entry can be classified into two groups: analytic and system implementation labels.

(10) pohon,ROOT+NOU+PS+TX+AP+ACS+ZN+DykaA1

3.1 Analytic label

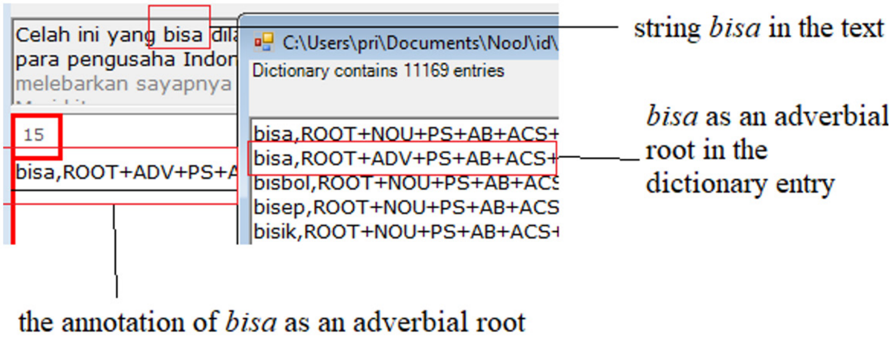
Analytic labels or Morphological Analysis (MA) labels are the reflections of formal and functional analytic categories users are likely to be interested in for searching. These labels are designed based on users' anticipated needs. For example, the monomorphemic entry *pohon* 'tree' has two analytic categories. The first label is ROOT, which signifies its formal category as a root morpheme. The second label is NOU, which corresponds to noun (the root's POS), a functional analytic category. They anticipate users query to search all instances of roots, or to search specifically all instances of noun roots. Following ROOT+NOU are system implementation labels, which will be discussed in the next section. In this section, all system implementation labels are omitted for easy-reading purposes.

(11) pohon,ROOT+NOU

The functional classification or roots is drawn from the common POS categorisation of Indonesian suggested by reference grammars of Indonesian (See Alwi et al. 1998; Sneddon et al. 2010). For example, *bisa* 'can/be able to' is an adverb of modality, and thus, is categorized as an adverbial root This differs from English, in which its equivalent, *can*, is likely to be classified as a modal verb or just modal. For instance, in CLAWS7 tag set (Garside 1987) the tag for can is VM (V=verb, M=modal), in which the modal is under the hierarchy of verb. Unlike CLAWS, in Penn Treebank tag set (Marcus et al. 1993), the tag for *can* is MD (modal), which is organized at the same hierarchy of verb tags.

Let us now get back to the adverb of modality, *bisa* in Indonesian. What analytic category is given to this root in SANTI-morf dictionary? While *bisa* includes the analysis of modality, only the highest hierarchy (adverb) is documented in the SANTI-morf tagset. Its specification as a modal is not given.

(12) bisa,ROOT+ADV



Celah ini yang bisa dila
para pengusaha Indon
melebarkan sayapnya

15

bisa,ROOT+ADV+PS+A

C:\Users\pri\Documents\NooJ\id\

Dictionary contains 11169 entries

bisa,ROOT+NOU+PS+AB+ACS+
bisa,ROOT+ADV+PS+AB+ACS+
bisbol,ROOT+NOU+PS+AB+ACS
bisep,ROOT+NOU+PS+AB+ACS
bisik,ROOT+NOU+PS+AB+ACS+

string *bisa* in the text
bisa as an adverbial root in the dictionary entry
the annotation of *bisa* as an adverbial root

Figure 2. *Bisa*: text, dictionary entry, and annotation

In Indonesian, the monomorphemic word *bisa* is actually ambiguous. It may refer to an adverbial root, as previously suggested or as a noun, which means ‘venom’. For ambiguous cases like this, the alternative analysis is also presented as a separate entry line. Thus, in addition to be analysed as an adverbial root (ROOT+ADV), *bisa* is also analysed as a noun root (ROOT+NOU). Ambiguities are resolved later using The Disambiguator module in SANTI-morf.

(13) bisa,ROOT+NOU

SANTI-morf analytic category labels also include *classifier*, a noun categorization morpheme (Sneddon et al. 2010:xxi). For instance, *ekor* is a classifier for nouns that fall under the category of animals. In Indonesian, this morpheme is bound to a numeral morpheme, thus, also called numeral classifier (Aikhenvald 2001:443). For instance, *ekor* in *dua ekor kucing* ‘two (animal classifier) cats’ is an animal classifier, as its occurrence is preceded by the numeral *dua* ‘two’. The classifiers modifies *kucing* ‘cat’, which is an animal. The majority of Indonesian classifiers are ambiguous. The morpheme *ekor* can also be used freely as a noun when it does not adjacently co-occur with numeral such as *ekor* in *ekor kucing* ‘cat tail’.

(14) ekor,ROOT+CLA (15) ekor,ROOT+NOU

Some root morphemes in Indonesian are bound; they cannot occur as a monomorphemic word. The root morpheme *juang* ‘struggle’ can serve to illustrate this. It can only occur within polymorphemic words such as *ber-juang* ‘struggle (intr)’, *per-juang-an* ‘struggle (noun)’, etc. The analytic category label +BOU is used to specify this root morpheme.

(16) juang,ROOT+BOU

There are 14 POS categories used as analytic category labels in SANTI-morf tagset. However, only 13 are true POS categories. The remaining one category is aimed to analyse foreign words, i.e., non-Indonesian words. Foreign words are analysed as monomorphemic even if in the source language they are monomorphemic. For example, ‘posting’ in English is polymorphemic. Regardless, SANTI-morf analyses it as a monomorphemic. For example, the word diposting is analysed as two morpheme tokens in which posting is treated as a root (ROOT+FRG).

Unlike MorphInd (Larasati et al. 2011), there is no ‘unknown’ POS category in SANTI- morf annotation scheme. When the Annotator module in SANTI-morf fails to perform an analysis, the Guesser will give its best guess rather than just leaving it unknown.

POS	Tag	Examples
Noun	NOU	*nasi* 'rice', *jagung* 'corn', London 'London'
Pronoun	PRO	*aku* 'I' (personal), *kenapa* 'why' (interrogative),*sini* 'this place' (demonstrative)
Numeral	NUM	*satu* 'one' (cardinal), *pertama* 'first' (ordinal)
Classifier	CLA	*ekor* 'animal class', *orang* 'human class'
Verb	VER	*pergi* 'go', *makan* 'eat', lari 'run'
Adjective	ADJ	*cantik* 'beautiful', *cepat* 'quick', *lama* 'long'
Adverb	ADV	*selalu* 'always', *jarang* 'seldom', *hanya* 'only'
Preposition	PRE	*di* 'at', *ke* 'to', *dari* 'from'
Conjunction	CON	*dan* 'and', *atau* 'or', *ketika* 'when'
Interjection	INT	*hai* 'hi', *aduh* 'ouch', *astaga* 'oh my god'
Article	ART	*si* 'the (derogatory)', *sang* 'the (honorific)'
Particle	PAR	*kah*, *lah*, *pun* (all emphasis)
Precategeorial	BOU	*juang* 'struggle', *nyanyi* 'sing'
Foreign	FRG	*post*, *posting* (English), *aqua* 'water' (Latin),*monggo* 'please' (Javanese)

Table 1. SANTI-morf root POS

In section 2, we discussed that a dictionary entry can be monomorphemic or polymorphemic. Earlier in section 3, we discussed the analytic labels and how they are formatted as a tag. However, the tag we just discussed only targets monomorphemic entry.

The tag format for polymorphemic entries slightly differs. This distinction is very important because the polymorphemic tag in the dictionary allows SANTI-morf to analyse the corresponding polymorphemic word in the text, exclusively using the dictionary, without having to work in unison with grammars.

A polymorphemic tag congregates the entry lines of all morphemes that form the polymorphemic word entry. They are presented in unison as a single line. Each is surrounded by angle brackets. For example, *tersangka* 'suspect (noun)' is a polymorphemic word composed of two morphemes: the agentive nominaliser teR- and the nominal root morpheme *sangka* 'suspect (verb)'. We will discuss the first morpheme, which is *teR-*.

The entry line for the first morpheme (prefix) is <ter,teR,PFX+R_NOU+AGNT+DykaA1>. The morpheme in this entry line has two forms, *ter* and *teR*. In SANTI-morf, this is a format given to a morpheme whose orthographic and citation forms differ. The presentation of both orthographic and citation form in the annotation output is required to anticipate users' need to carry out both specified and underspecified searches.

For instance, morpheme *teR*- has two allomorphs, *te*- and *ter*-. In some cases, a user might want to retrieve word forms containing either *te*, or *ter*-. However, in some cases, a user might want to retrieve word forms containing both *te*-, and *ter*-. In this case, the user will only need to query with <teR> to get both *ter*- and *te*.

This approach applies to the overall annotation output, either generated by dictionaries, grammars, or their combinations. So, for instance, the active verb *meN*- has 6 allomorphs, whose annotation output is created using grammars, not dictionaries. To get all word forms containing these allomorphs, users will just need to query with the orthographic form <meN> and get all word forms containing the allomorphs.

Let us discuss the corresponding analytic labels for ter and teR. PFX refers to prefix. R_NOU refers to the outcome POS of the prefix, here to derive a noun. PTNT refers to a patient. Overall, these labels signify a patientive nominaliser tag. These tags are usually incorporated into morphemes, which are written in the grammar not the dictionary. As we are not focusing on the grammar, I will not discuss other labels further. DykaA1 is an implementation label, not an analytic. This label will be discussed in the next section.

Now, let us get back to the sequence of morphemes that form *tersangka*. Following the agentive nominaliser prefix *ter*- is a verbal root morpheme *sangka* 'to suspect', whose entry line is <sangka,ROOT+VER+DykaA1>. Now that all the required entry lines from the two morphemes are available, we need to concatenate them in the tag slot as a single entry line. The full word entry line is shown as follows.

(17) tersangka,<ter,teR,

PFX+R NOU+PTNT+DykaA1><sangka,ROOT+VER+DykaA1>+UNAMB

This entry line allows SANTI-morf to annotate *tersangka*, without having to combine it with any grammar. While this applies to the specific polymorphemic word *tersangka*, it does not apply to all words even in the same word family. For example, the polymorphemic word *disangka* 'to be suspected' and *menyangka* 'to suspect' are not solely annotated using the dictionary even though they share the same verbal root *sangka*.

The two words are produced using productive morphemes meN- (active verb prefix) and di- (passive verb prefix). These two morphemes are incorporated into SANTI-morf grammars, not dictionaries. To analyse *menyangka* and *disangka*, the grammars need to work together with dictionaries to analyse polymorphemic words. The tag ends with +UNAMB. This is a system implementation label. Thus, it will be discussed in the subsequent section.

3.2 System implementation label

System implementation labels are labels used for SANTI-morf implementation purpose. For example, in all annotation outcomes given by SANTI-morf, we can observe the name of the source file which generates the annotation at the very end of each tag. For dictionaries, there are only three possible labels from three dictionary files, arbitrarily named as follows: DykaA1, DykaA2, and DykaA3.

DykaA1 consists of common entries, which are not proper names or foreign word. For instance, *pohon* 'tree' is one of the entries in DykaA2 as it is not a proper name or a foreign word. DykaA2 consists of proper name entries such as *Aljazair* 'Algeria'. DykaA3 consists of non-Indonesian entry such as *response* (from English).

(18) pohon,ROOT+NOU+DykaA1

(19) Aljazair,ROOT+NOU+DykaA2

(20) response,ROOT+FRG+DykaA3

The label of the name of the resource file can be used for debugging purpose. For instance, when an error is detected in the annotation outcome, a developer can quickly retrieve the resource file suspected to be the source of the error. S/he then can locate the specific entry line and modify it.

In addition to resource file name, system implementation labels also include labels used for rule constraint purpose. Orthography labels can serve to illustrate this. For instance, the verbal root morpheme *cari* 'search' ends in i, and thus is marked with +ZI label. One of the affixation rules in Indonesian reference grammars (Alwi et al., 1998:117) dictates that suffix - i cannot attach to bases ending in i. SANTI-morf takes this rule into account. Once this label (+ZI) is detected, the *-i* suffixation rule is blocked for the corresponding root entry (*cari, lari, beri* and all root morphemes ending in *-i*). Other labels are given in table 2.

(21) cari,ROOT+VER+PS+AC**+ZI**+ACS+T1+DykaA1

Let us consider another example, this time from syllable number label. The root entry *las* 'wield' is a monosyllabic root entry, thus, consists of +MS label. This is a useful label for selecting the correct allomorphs. For example, *meN*- and *peN*- morphemes have six allomorphs each. However, when the corresponding root is monomorphemic the correct allomorphs are *menge*- and *penge*-.

(22) bom,ROOT+VER+**MS**+AB+ZI+ACS+TX+DykaA1

The last category in the system implementation labels is transitivity. Each verb root entry are marked for their transitivity, either they are intransitive (+T0), transitive (+T1) or ambitransitive (+T2). Non-verbal root entries are given +TX label. Transitivity label is actually a grey area label between analytic and system implementation labels. I use this to set constraints. For example, the reciprocal function for circumfix *ber—an* is added into the annotation when the verb root is transitive (*tabrak* 'hit (trans)' > *ber-tabrak-an* 'hit one and each other' VS *jatuh* 'fall (intr)' > *ber-jatuh-an* 'fall randomly'). While I use this label for implementation purpose, nothing actually prevents users from using it to retrieve, for example, all words whose roots are transitive.

(23) tabrak,ROOT+VER+PS+AT+ZK+ACS+**T1**+DykaA1

System implementation labels	

Dictionary name	DykaA1 = main dictionary
	DykaA2 = proper name dictionary
	DykaA3 = foreign word dictionary

Syllable	MS = Monosyllable
	PS = Polysyllable

Orthography	AA = begins with letter a AB = begins with letter b … ZA = ends with letter a ZB = ends with letter b … AVW = begins with vowel ACS = begins with consonant

Transitivity	TX= non-verb
	T0 = intransitive
	T1 = transitive
	T2 = ambitransitive

Table 2. System implementation labels (… : omitted labels due to space constraint)

A special label +UNAMB shown at the end of section 3.2, finalises the tag which corresponds to the word form entry *tersangka.* This is one way to perform disambiguation, which can be illustrated as follows.

In the grammar, there are some rules with *ter-* prefix. Without +UNAMB label in the corresponding entry line for *tersangka* in the lexicon, SANTI-morf would generate the annotation given by the lexicon as well as by the rules in the grammar files. This means, there would be multiple annotations on the same words (i.e. ambiguity). However, with the special label +UNAMB given to finalise the corresponding tag for *tersangka* in the lexicon, all the annotations from the rules are blocked. Thus, only the annotation from the lexicon, *ter-* as an patientive nominaliser, is produced. It then overrides the analyses of *ter-* as a verbal prefix.

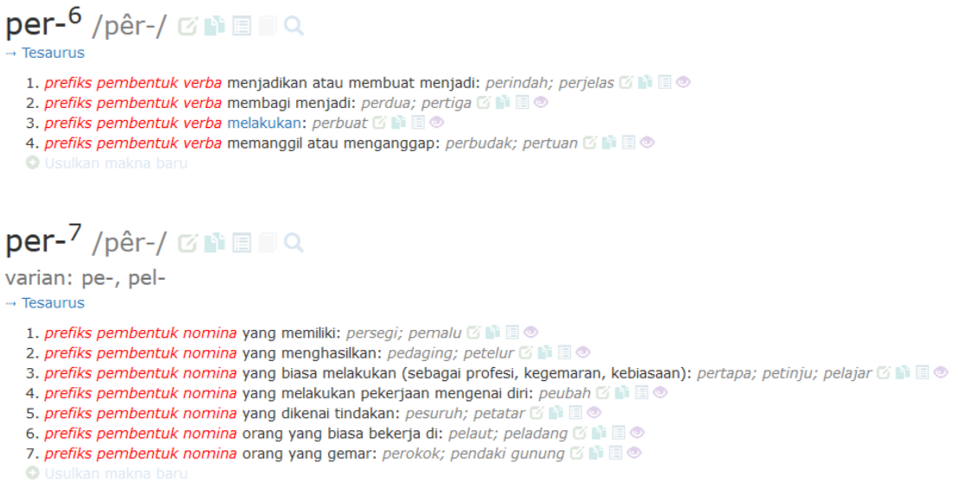
3.3 Residual label

A residual label is a term assigned to non-letter characters. These non-letter characters are grouped into two categories: numeric digits (DGT) and punctuations (PUNC). However, only punctuations[7] are listed as entry lines in the dictionary. In the entry line, every punctuation is identically tagged, using only one label PUNC, followed by the file name.

(24) :,PUNC+DykaA1

4. Morpheme list and frequency information

SANTI-morf can be used for a variety of applications. However, let us now focus on using SANTI-morf to support lexicographic work, particularly on supplying frequency information. Consider the description of *per-* entry obtained from the online[8] *Kamus Besar Bahasa Indonesia* (KBBI), or in English, the Great Dictionary of Indonesian.



per-^{6} /pêr-/

→ Tesaurus

1. *prefiks pembentuk verba* menjadikan atau membuat menjadi: *perindah; perjelas*
2. *prefiks pembentuk verba* membagi menjadi: *perdua; pertiga*
3. *prefiks pembentuk verba* melakukan: *perbuat*
4. *prefiks pembentuk verba* memanggil atau menganggap: *perbudak; pertuan*

Usulkan makna baru

per-^{7} /pêr-/

varian: pe-, pel-

→ Tesaurus

1. *prefiks pembentuk nomina* yang memiliki: *persegi; pemalu*
2. *prefiks pembentuk nomina* yang menghasilkan: *pedaging; petelur*
3. *prefiks pembentuk nomina* yang biasa melakukan (sebagai profesi, kegemaran, kebiasaan): *pertapa; petinju; pelajar*
4. *prefiks pembentuk nomina* yang melakukan pekerjaan mengenai diri: *peubah*
5. *prefiks pembentuk nomina* yang dikenai tindakan: *pesuruh; petatar*
6. *prefiks pembentuk nomina* orang yang biasa bekerja di: *pelaut; peladang*
7. *prefiks pembentuk nomina* orang yang gemar: *perokok; pendaki gunung*

Usulkan makna baru

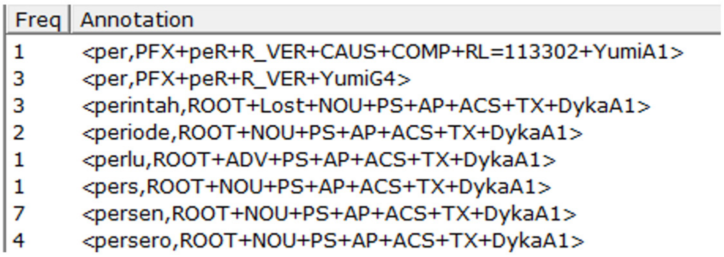
Figure 3. KBBI description for per- entry

A search with *per-* query returns two entries: *per-* whose outcome POS is a verb (per- ^{6}) and noun (per-^{7}). The senses for these two entries vary, but none has frequency information. In fact, frequency information is a feature absent in all KBBI entries. A frequency information that corresponds to an entry, can be automatically derived from a corpus. However, for bound entry like *per-* the corpus must have been annotated at morpheme level.

SANTI-morf carries out annotation at morpheme-level. Thus, it can produce a morpheme list. The morpheme list in SANTI-morf includes frequency information, as shown below.

[7] It is more effective to annotate numeric digits using grammars rather than dictionaries as they have too many combinations.

[8] https://kbbi.kemdikbud.go.id/



Freq	Annotation
1	<per,PFX+peR+R_VER+CAUS+COMP+RL=113302+YumiA1>
3	<per,PFX+peR+R_VER+YumiG4>
3	<perintah,ROOT+Lost+NOU+PS+AP+ACS+TX+DykaA1>
2	<periode,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
1	<perlu,ROOT+ADV+PS+AP+ACS+TX+DykaA1>
1	<pers,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
7	<persen,ROOT+NOU+PS+AP+ACS+TX+DykaA1>
4	<persero,ROOT+NOU+PS+AP+ACS+TX+DykaA1>

Figure 4. SANTI-morf morpheme list and frequency information sample

We see that there are two *per*- in the morpheme list. The first item has only one instance and the second item has 3 instances. While the orthographic forms are identic, the tag differs. For this reason, SANTI-morf presents them as two separate items in the morpheme list. Note that both contain R_VER analytic labels. This means, the outcome POS for these instances is all verbs, which corresponds to KBBI entry of per-^{6}.

The item which contains the +CAUS analytic label corresponds to the first sense in *per*-^{6}, whose frequency is only one[9]. The second item, whose frequency is 3, does not contain +CAUS. It corresponds to the remaining of the senses (2,3, and 4)[10].

Note that the frequency information in this paper is obtained from a small corpus, thus may not be representative of the Indonesian language. With a larger corpus, more representative frequency information can be obtained. The frequency information can be linked to each sense, allowing KBBI to produce frequency information automatically. This frequency information can enrich KKBI entries description.

5. Conclusion

In this paper, I have described the architecture of SANTI-morf dictionaries. These dictionaries work together with other SANTI-morf components allowing SANTI-morf to automatically annotate Indonesian texts. I have also demonstrated the application of SANTI- morf, in this case, by supplying frequency information for the bound morpheme entry *per*- in KBBI. While additional mechanisms are required to port SANTI-morf to KBBI, including creating a corpus from which frequency information can be automatically derived for each entry, such demonstration illustrates how SANTI-morf can be used to support lexicographic works, and potentially works across disciplines, such as in corpus linguistics, informatics, etc.

References

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, M. (1998). *Tata Bahasa Baku Bahasa Indonesia (3rd Edition)*. Jakarta: Balai Pustaka.

Larasati, S.-D., Kuboň, V., & Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Systems and Frameworks for Computational Morphology* (pp. 119-129). Zurich, Switzerland: Springer Berlin Heidelberg.

[9] This low frequency is likely to be caused by the size of the corpus, which indeed is small, as used for presentation purpose only.

In the morpheme list, there is no item which corresponds to *per*-^{7} whose outcome POS is noun. This is also likely to be caused by the small corpus size. For more serious works, a large corpus data need to be collected and analysed using SANTI-morf.

[10] At present, SANTI-morf has not been imbued with semantic information, thus cannot distinguish senses number 2,3, and 4.

Lewis, M.-P., Simons, G.-F., & Fennig, C.-D. (2009). *Ethnologue: Languages of the world (Vol. 16).* Dallas: SIL International.

Marcus, M.-P., Marcinkiewicz, M.-A., & Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics 19(2)*, 313–330.

Pisceldo, F., Mahendra, R., Manurung, R., & Arka, I.-W. (2008). A Two Level Morphological Analyser for the Indonesian Language. *Australasia Technology Association Workshop*, (pp. 142-150).

Prihantoro. (2019). A new tagset for morphological analysis of Indonesian. *International corpus linguistics conference.* Cardiff.

Prihantoro. (2021). An Evaluation of the Morphological Annotation Scheme for Indonesian Used in MorphInd Program. Corpora, in Press . *Corpora 16 (3)*, in press.

Prihantoro. (2021). *SANTI-morf: a new morphological annotation system for Indonesian (a PhD thesis:forthcoming).* Lancaster: Lancaster University Press.

Silberztein, M. (2003). *NooJ Manual.* Available for download at: www.nooj4nlp.net. Silberztein, M. (2016). *Formalizing Natural Languages: Nooj Approach.* London: Wiley. Sneddon, J.-N., Adelaar, A., Djenar, D.-N., & Ewing, M.-C. (2010). *Indonesian reference grammar:2nd Edition.* New South Wales: Allen & Unwin.

PROBLEMS AND CHALLENGES IN WRITING GERMAN-INDONESIAN PHRASEOLOGICAL LEARNER'S DICTIONARY

Raden Muhammad Arie Andhiko Ajie
Universitas Indonesia
immershend7@gmail.com

Abstract

Writing a phraseological dictionary is a challenging and time-consuming work. The challenge increases when there is no adequate sources of examples of phraseological use and existing phraseological dictionaries in target language. Prior to the writing process, lexicographer should conduct a needs assessment to determine the macrostructure and microstructure of the dictionary. Needs assessment, which is crucial to make a feasible plan, includes a comprehensive observation of the existing phraseological dictionaries in source and target language, as well as identifying potential users of the dictionary. This is a pivotal stage because economic aspect is very important in publishing a dictionary.

This paper discusses problems in writing German-Indonesian phraseological dictionary, specifically the problem in determining the German lexical entries; in explaining the phraseological meaning; in dealing with the lack of up-to-date Indonesian phraseological dictionaries and in finding equivalences for the phraseological expressions/ phrasemes; as well as in finding and modifying the appropriate examples of using the German phraseological expressions; and in overcoming other practical problems. This paper aims to show various solutions to the problems based on my experience in writing German-Indonesian phraseological dictionary, like using Internet as a form of validation to the lexical entries. Moreover, it also aims to explain simple steps to make a phraseological dictionary a learner's dictionary.

Keywords: phraseological dictionary, phraseological expressions, phraseological equivalences, corpus validation, German-Indonesian dictionary

1. Introduction

Writing an ordinary dictionary that suits the objectives and standards of lexicology and lexicography takes a long time and persistence because we need to consider many factors. These factors will manifold when we write a phraseology dictionary. One of the essential factors is to understand the characteristic of phrasemes to make sure that the dictionary in the making will show its particularity.

To distinguish phraseology from a series of words, we need to study carefully existing phraseology books. One of the recommended phraseology books is by Burger (2015). This book has been printed and published for the fifth time since its first edition and closely observes the writer's role in the development of phraseology, especially in German linguistics. That is, in my opinion, a solid reason to choose Burger (2015) as a reference in understanding phraseology, especially phraseology in German.

A dictionary is a potential resource in dealing with phraseme. An excellent general monolingual or bilingual dictionary should assist in terms of translation or oral communication. Foreign language teachers, foreign language learners, translators and those confronted with German phraseological expressions or phrasemes need phraseological dictionaries to refer to when understanding or producing phraseme. However, usually, the number of dictionary articles in a general dictionary that takes phraseme into account is minimal.

There is already a dictionary of Indonesian-German phraseology, namely *Andere Wiese, andere*

Grashüpfer ... oder Andere Länder, andere Sitten. Indonesisch-deutsche Sprichwörter und Redensarten im Vergleich (2015) written by Herlina and Nandzik. This dictionary contains many Indonesian expressions and proverbs, as well as their German equivalent. After the phraseological form (lemmata), the dictionary provides a literal translation of the components of the Indonesian phrasemes, followed by an explanation of the meaning of the phraseological meaning in German and the equivalent in the German phraseme, if available. However, this dictionary does not provide phrasemes variations, description of the structure of phrases, stylistic descriptions of phrasemes and examples of their use. Moreover, many of the entries in the form of phrasemes are unpopular and seem outdated. Therefore, I find this dictionary to be not suitable to be a learner's dictionary. The target users of this dictionary are German-speaking people, which means that there is still free space for a phraseological dictionary targeting German learners as its potential users.

Based on the inadequacy of the German-Indonesian phraseological dictionary, I find it essential to write a learner's dictionary that meets the requirements of lexicology and lexicography. This dictionary can support the process of learning German because it provides more than just the meaning of the phraseme. It will also include examples of using phrasemes in communication, which will help learners understand the meaning easily and learn how to use phrasemes actively.

This article aims to provide an overview of making bilingual phraseological dictionaries for learners. I hope that explaining the factors that we need to consider and the necessary steps can contribute to lexicography in Indonesia. The explanation in this article is certainly not the most ideal and complete steps in making dictionaries, but it can serve as an alternative in designing a bilingual phraseology dictionary for foreign language learners.

This article highlights the complexity of writing a German phraseology dictionary for Indonesian native speaker. The first challenge in this process is to connect German phraseology with Indonesian because phraseology is not yet established as a linguistics branch. In Indonesian linguistics, the term phraseology, which includes expressions and proverbs, is still unfamiliar, and people usually refer to the dictionaries of Indonesian phrasemes as idiomatic dictionaries. Some proverb dictionaries also include other content, such as enhanced spelling rules (*Ejaan Yang Disempurnakan* or EYD). This is a common practice because proverb dictionaries are not exactly best-sellers.

The second challenge is to find existing dictionaries. The latest good-quality dictionary, which contains Indonesian expressions, is Mahayana, Nuradji and Totok Suhardiyanto, published in 1997. The period of more than two decades allows the possibility of exclusion of many recent expressions. We should be mindful of this factor because language is constantly evolving, and some expressions are getting outdated and gradually unused. This is a crucial factor because the purpose of making a dictionary is not just to document words but also to create a dictionary for learners who actively use the language. Therefore, we need to filter the entries for the dictionary. We can use the corpus linguistic method to select the entries.

The lack of new, documented expressions and the microstructure of a simple Indonesian dictionary makes it difficult for lexicographers to find the correct equivalent in Indonesian. Several dictionaries of Indonesian expressions show deficiencies in explaining the meaning, and there are no examples of usage. Inconsistency is also common: some dictionaries contain examples, and some others do not. There are even dictionaries that provide examples only at the beginning of the dictionary section. This inconsistency shows that writing a dictionary does require persistence and a long time. Like marathon runners, we need high stamina to maintain the quality of the dictionary from start to finish.

Before the writing process, the lexicographer certainly needs to think about macrostructure and microstructure. The macrostructure of the dictionary concerns the entries selection and their order, while the microstructure corresponds with the dictionary elements. The lexicographer needs to identify the potential users of the dictionary to know their need for a dictionary. Creating a learner's dictionary is more challenging than a general dictionary. We need to carefully plan the number of entries, variations of entries and exercises. A good plan is crucial to ensure the quality and selling point of the dictionary.

As mentioned earlier, this article explores the writing process of a German phraseological dictionary for Indonesian. Therefore, the potential users of this dictionary are German learners whose mother tongue is Indonesian. Indonesian native learners who learn German will undoubtedly benefit from learning German phrasemes to expand their vocabulary and communicate with better precision. This dictionary aims to be a bridge for learners to learn German phrasemes quickly, precisely, and clearly. In addition, this dictionary can act as a bridge for elementary-level learners because it is bilingual. A monolingual phraseological dictionary such as Duden Redewendung Dictionary of German Idiomatics 4th revised and updated edition, or the Duden Volume 11 (henceforth: Duden 11), offers much more extensive and better explanations than information in standard dictionaries. However, as a monolingual dictionary, Duden 11 can be too challenging for German learners at a basic level. In other words, a bilingual phraseological dictionary is more suitable than a monolingual dictionary for those who only have limited German knowledge.

This dictionary is designed to contain phrasemes commonly used in modern German. The number of entries is limited to 100-200 phrasemes. In selecting the phrasemes, I consider two important factors. The first factor is the usage frequency of the phrasemes, and the second is their equivalence, namely the existence of Indonesian phrasemes that correspond to the German phrasemes. The limited number of phrasemes is closely related to the purpose of making the dictionary. A smaller number of entries will increase the learner’s focus in learning the use of the phraseme. In addition to this, it is also more feasible for lexicographers to write highly qualified dictionaries.

2. Method

This research is qualitative research based on experience in compiling a German phraseology dictionary. The compilation of this dictionary was made possible by studying the literature of various works from German language phraseology, lexicology and lexicography.

It is essential to analyse the potential users of a dictionary before designing the dictionary. This phase will help us prepare the following steps: designing the macrostructure and microstructure of the dictionary, determining the entries and their quantity, and understanding and explaining the restrictions of German phraseological expressions. To design a suitable microstructure, we can use Duden 11 Dictionary as a guide. Duden 11 is known as the largest and most general dictionary of German expressions. The Duden 11 microstructure is already very good. However, this dictionary is not suitable for German learner at the beginner’s level.

This article also discusses the problems emerging in explaining the phraseological meaning, in dealing with the lack of up-to-date Indonesian phraseological dictionaries and finding equivalences for the phraseological expressions, and in finding and modifying the appropriate examples of using the German phraseological expressions. The next part of this article discusses the role of corpus linguistics. This method has also become an inseparable part of the compilation of dictionaries today because it is the scientific way to validate the usage and the familiarity of the phraseological expressions. Finally, this article also explains the forms of exercise that can be applied. This exercise part is the one that makes a phraseological dictionary can be called a learner’s dictionary.

3. Result

There are many challenges in compiling a learner’s bilingual phraseology dictionary. Some of the challenges are in the following areas:

1. in designing the microstructure and the macrostructure of the dictionary;
2. in analysing the syntactic restrictions and the vocabulary restriction;
3. in explaining the phraseological meaning;
4. in determining the German lexical entries;
5. in finding and modifying the appropriate examples of using the German phraseological expressions;
6. in finding equivalences for the phraseological expressions in the Indonesian language, especially in finding the actual equivalence.

Solutions:

1. To concept the microstructure and the macrostructure of the dictionary effectively, we have to observe and study similar dictionaries. First, we should try to analyse their strength and weakness. The results can help us design the microstructure and the macrostructure of the dictionary in a relatively short time. Then, of course, we have to concept the microstructure and the macrostructure after analysing the needs and the behaviour of the potential dictionary users.
2. To analyse the syntactic restrictions and the vocabulary restrictions of German phrasemes, we need to study research articles from phraseology. Just by studying the information in Duden 11, sometimes it is not clear enough. It needs a deep understanding of the German language to be able to recognise these restrictions. In some cases, it is also crucial to discuss some unanswered questions with native German speakers with excellent language knowledge.

   The dictionary writer can take the form of the entry listed in Duden 11 and clarify it when it is necessary. For example, they can provide additional information, whether the subject or object of a phraseme can be people or things. If the subject or object is a person, we must observe it further to determine whether the phraseme position should be filled with a specific gender.
3. To explain the phraseological meaning, I recommend taking the explanation from a complete and tested dictionary, namely Duden 11. If the explanation of the meaning in Duden 11 is somehow inadequate, this dictionary consistently provides authentic examples. An authentic example of using a phraseme can help to explain the meaning of the phrase. The language used to explain the meaning should be the native language of the dictionary user. Since the target dictionary users are German learners from a basic level, it is best to write the explanation in Indonesian.

   If the explanation of the meaning and examples in Duden 11 is not enough, the lexicographer can conduct online studies. Other descriptions from various internet sources are helpful in this regard. If it is still considered unsatisfactory, the lexicographer should discuss it with someone whose mother tongue is German.
4. For creating a learner's dictionary, the number of entries cannot be as many as possible. Here it is necessary to select the frequency of using a phraseme. The corpus linguistics method is a way out to help select a phraseme in everyday communication. In the initial steps of determining which phrasemes to use, dictionary compilers can do a short study to find out the frequently used phrasemes. This study can be done by looking at a list of frequently used phrasemes or marking the frequently used phrasemes in a dictionary. If there are too many collected phrasemes to be included in the dictionary, the dictionary-maker can select them by examining their frequency of use using the corpus linguistics method. Using the language of several sources can also be seen as an alternative.

   On the contrary, if the number of phrasemes collected is still less than the target, lexicographers can add it by searching again in other phraseological dictionaries. Doing an online study on online media is also good because many new phrasemes have not been documented in the dictionary. Thus, we can use the internet to validate whether a phrase is indeed used and known in society. The next step is to ask for help from sources who speak the target language. Their opinions can serve as validation for the information that the lexicographers have already obtained.
5. The dictionary should provide authentic examples to explain how to use a phraseme. Duden 11 consistently provides authentic examples of the use of phrasemes to make it easier for readers to see the structure of the phrasemes and better understand the meaning of phrasemes. However, many authentic examples are difficult to understand without extensive knowledge. Another problem, there are authentic examples that are pretty long, and the structure is difficult for elementary level learners. Therefore, the example made by the lexicographers can be a solution because it can be made simpler and shorter. The lexicographers do not need to create new examples but can modify existing authentic examples to simplify them. For example, we can give a changed subject and object or sentence structure as an example. However, here we need sources who can determine whether the examples are acceptable and natural.

6. Due to the lack of Indonesian phraseological dictionaries, it is difficult to rely on printed dictionaries alone to look up German equivalent phrasemes in Indonesian phrasemes. The lexicographers need to do online studies and rely on their knowledge of the language to find the exact equivalent. The positive side is, the lexicographers can find more actual equivalents. This online study requires attentiveness, and the lexicographers should use valid and relevant online sources. It would be better if the sources are online mass media.

   Determining phrasemes equivalents is not easy. First, of course, meaning is the main factor in determining equivalents. However, the form or component of the phrasemes can also be considered when looking for equivalents. Another thing to consider is the frequency of use of equivalent phrasemes. It may be that these phrasemes have the same meaning and form, but they are unfamiliar, while some others are more commonly known. In this case, it is essential to prioritise the more commonly used phrasemes so that learners can use them in communication.

4. Analysis and Discussion

Concepting the microstructure and the macrostructure of the dictionary

In making a dictionary, lexicographers must analyse the needs of their users and design suitable dictionaries based on their needs. In this case, the macrostructure and microstructure must be well thought out. For example, what lemmata are documented, the number of lemmata, the explanation of the meanings, any description of language variety, any examples, and whether there should be an internal structure that links one entry to another in the dictionary.

To determine the macrostructure and microstructure of a dictionary effectively, the lexicographer can study an existing dictionary. The ideal course is to study the existing phraseological dictionaries. The dictionary should be able to provide a brief description of the phrasemes. According to Burger (2015), there are two criteria to determine whether a word bundle is a phraseme: 1. Polilexicality/*Polilexikalität* (a phraseme consists of at least two words), 2. Firmness/*Festigkeit* (This criteria means when the phraseme is used in precisely this (or a very similar) combination of words. Speakers in a language community use phrasemes similar to a word. The criteria of firmness are the most complicated compared to two other mentioned properties of phrasemes because they can be analysed from the perspective of usage, structure, pragmatism, and style. Finally, there is the third criteria, and if this last criteria fulfiled, the phraseme is an idiom. The third criteria is idiomaticity/*Idiomatizität*. A phraseme is idiomatic if we can not derive its phraseological meaning from the application of syntactic and semantic rules.

The microstructure of a dictionary includes the content and structure of a dictionary article (Schaeder 1987). Microstructure includes all the contents of the explanation of the entry, the functions, and punctuation in the dictionary. In designing the German-Indonesian phraseological dictionary, it is necessary to pay attention to the writing of entries. Duden 11 can serve as a basic example, but it is still incomplete. Because the reader is a foreigner, it needs to be clarified whether the subject or object in the phraseme is a human or an object. It should also be explained whether there is a grammatical restrictions in the phraseme. If the phraseme is only in the form of the perfect tense, of course, it is written differently. For example, *einen Narren an jmdm. gefressen haben* instead of *einen Narren an jmdm. fressen* because we can use this phraseme only in the perfect tense.

In addition, it is necessary to think about the explanation of the meaning of the phraseme, its equivalent in Indonesian, explanations of grammar and pragmatics, examples of use, the origin of the phraseme and other information, if any.

In constructing the microstructure, Duden 11 can provide guidance and provide a bunch of information. However, what needs to be adjusted are the examples of the usage of the phrasemes. The examples in Duden 11 are authentic. However, some examples are difficult to understand because of complex sentence construction, which requires additional knowledge. In this case, the lexicographer should be able to make natural and straightforward examples. In giving these examples, the lexicographer can use the internet or ask native German.

The following are examples of suggested dictionary elements for a German-Indonesian phraseological dictionary for learners.

1. Lemma as the first element, this lemma is equipped with grammatical and stylistic explanations (in German)
2. Explanation of the meaning of the phraseme (in Indonesian)
3. Equivalent (If there is a total equivalent)
4. Usage examples (in German)
5. Additional information on the origin of the phraseme (in Indonesian)
6. Instructions that direct dictionary users to look at phrasemes that have related meaning, whether they have similar meanings, have similar meanings or have opposite meanings

Analysing the syntactic restrictions and the vocabulary restrictions

In the use of phraseme, learners must understand the restrictions. Burger (2015) mentions two types of restrictions, namely morphosyntactic restriction (*Morfosyntaktische Restriktionen*) and lexical-semantic restrictions (*lexikalisch-semantische Restriktionen*). The problem is, Duden 11, as a source, does not explicitly mention these restrictions. For example, certain phrasemes are only usable in the perfect form. There are also certain phrasemes where a specific subject or object can only fill the subject or object. Restrictions are written in lemmata and can also be identified in the examples given; of course, we need good language knowledge in this case. In making a lemma, known as the lemmatisation process, lexicographers must pay attention to the morphosyntactic and lexical-semantic restrictions of a phraseme (Burger 2015).

We cannot change the phraseme "*das ist kalter Kaffee*" into one main sentence with one clause, namely "*Das ist Kaffee, der kalt ist*" or using the plural form "*Das sind kalte Kaffees*". If this restriction is violated, this expression can lose its phraseological meaning.

Certain phrasemes are tied to tenses, for example "*einen Narren an jemandem gefressen haben*". In Duden 11, this restriction can be seen from the lemma used, namely "*einen Narren an jemandem gefressen haben*" instead of "*einen Narren an jemandem fressen*":

Romeo frisst einen Narren an Julia. (incorrect)
Romeo fraß einen Narren an Julia. (incorrect)
Romeo hat einen Narren an Julia gefressen. (correct).

Another restriction described by Burger (2015) is the lexical-semantic restriction. Thus, for example, the phraseme "*die Flinte ins Korn werfen*" cannot be changed to "*das Gewehr ins Korn werfen*", even though the words *Flint* and *Gewehr* have similar meanings.

Römer and Matzke (2005) argue that other restrictions are related to the subject, namely gender restrictions. There are subjects whose subjects can be filled by women or men, but the meanings can be different, namely:

a. *Sie kam in voller Kriegsbemalung*. (She came in full-body painting) "She wore eye-catching makeup."
b. *Er kam in voller Kriegsbemalung*. (He came in full-body painting) "He came with all his medals and awards."

Some phrasemes have non-idiomatic meaning if the subject is a man. For example:

a. *Sie hat viel Holz vor der Hütte*. [idiomatic] "She has big breasts".
b. *Er hat viel Holz vor der Hütte*. [not idiomatic] "He has many planks of wood in front of his cabin".

These examples show us the importance of understanding the existence of these restrictions. This information is needed to instruct the dictionary user to use the phraseme correctly. According to Kreuder (2003), information about these restrictions is essential for novice learners. This information is a positive indication that a lexicographer pays attention to his dictionary users.

The solution and the entry or lemmata must be able to show the existence of these restrictions. Grammatical explanations are also needed to make it easier for dictionary users to recognise the restrictions of a phraseme. In addition, in the exercise section, these restrictions can be discussed again; for example, we should add the information about the morphosyntax or lexical-semantic restrictions to the answer key.

Explaining the phraseological meaning

Duden 11 includes an excellent and user-friendly section, which explains the phraseological meaning. This dictionary was developed by a very competent team and has been through five updates. I recommend Duden 11 for those who intend to understand the phraseological meaning of German Phrasemes because it provides excellent explanations about the phraseological meaning. This dictionary also includes authentic examples to help the users understand the usage and the meaning of phrasemes in communication. Furthermore, the explanation in a phraseological dictionary for learners is ideally in the native language of the dictionary user. Therefore, I choose to write the phraseological meaning in Indonesian because the target users are German learners in the beginner-level.

If the explanation of the meaning and examples in Duden 11 is not enough, we can conduct online studies. Other descriptions from various internet sources are helpful in this regard. If it is still considered unsatisfactory, I suggest discussing it with a German native speaker with excellent language competency.

According to Lemnitzer and Zinsmeister (2015), corpora serve as an important source for usage examples. Compared to the linguistic competence of lexicographers, corpora are superior. The ability to show the frequency of occurrence of certain lexical units is why corpora can help in the selection of lemmata. This possibility is more important, especially when it comes to creating a learner's dictionary. Individual decisions of the lexicographers can be checked against corpora not only up to the production phase of a dictionary, but also in the correction phase.

There are limits to using corpora in a corpus linguistic study. Corpus size is not always satisfactory, especially when it comes to identifying idiomatic phrasemes. The use of phrasemes, especially idiomatic verbal phrasemes, is very low in various languages (Colson 2007). This is the reason, why lexicographers using internet as a method to validate the lemmata or the examples of the phrasemes used in authentic communication.

The next step for creating a learner's dictionary is creating exercises. The exercises can be started from recognizing the phrasemes, remembering the component of the phrasemes, understanding the meaning of the phraseme and finally using the phraseme. Exercises that strengthen the understanding of the meaning of the phrasemes can be in the form of multiple choice or matching exercises. On the one hand, the phraseme is displayed, on the other hand, the exact meaning of the phraseme and its distractor is displayed. The phraseological dictionary which is a learner's dictionary that can be used as a reference in making exercises is the *Portugiesische Redewendungen. Ein Wörter- und* Ü*bungsbuch für Fortgeschrittene* by Ettinger und Nunes (2006). They made a bilingual phraseological dictionary, namely Portuguese-German, and the exercises.

Determining the German lexical entries

According to Lemnitzer and Zinsmeister (2015), corpora can be used in dictionary planning right from the start in order to calculate the object to be described by the dictionary. Corpora can provide hints for choosing a lemma, so that the frequency of a lexical unit can be determined quickly. This makes it clear whether or not this lexical unit should be included in a dictionary. Corpora also serve as a source of information when creating a dictionary, since they contain various information on lexical units at all linguistic levels. It is the task of lexicographers to analyze and filter out the necessary information on certain lexical units for the dictionary article.

For the creation of a learner's dictionary, the number of entries cannot be as many as possible. It is necessary to select the frequency of using a phraseme. The corpus linguistic method is a way out to help select the use of a phraseme in everyday communication. In the initial steps of determining which phrasemes to use, lexicographers can do a short study to find out what phrasemes are often used. Using corpus linguistic methods is known as the corpus lexicography (Engelberg/Lemnitzer 2009).

Before doing the selection of the phraseme, lexicographers have to collect first so many phraseme from various resources. Using the language of several resources can also be seen as an alternative. If the number of phrasemes collected is still less than the target, lexicographers can add it by searching again in other phraseological dictionaries. Doing your own online study on online media is also good because there are many new phrasemes that have not been documented in dictionaries.

After the phrasemes that want to be presented in the dictionary are collected and considered too many, lexicographers can select them by examining their frequency of use using computer and internet. Thanks to this technology, lexicographer can validate the language data they get. They can make sure, if the phraseme actually used and well known, not phraseme that already obsolete.

The next step that can be taken is to ask for help from sources who speak the target language. Their opinions can serve as validation for the information that the lexicographers have already obtained.

The selected phrasemes can be based on the criteria for frequency of use. Another factor that can be taken into account is whether there is an equivalent in the mother tongue of the dictionary user. Good literature to understand about equivalent phrasemes and their classification is the articles from Koller (2007) and Korhonen (2007).

In making a phraseological dictionary for learners, it's a good idea to look for German phrasemes that have the same meaning as phrasemes in Indonesian, for example "*jemand ist noch grün*" (someone is still green) and "*die Katze im Sack kaufen*" (to buy a cat in a sack), became the choice of lexicographers.

Phrasemes that are also interesting to choose are what are called internationalism, namely phrasemes that are found in various languages. Although there are slight structural or lexical differences, the phraseological meaning remains the same, for example "*Viele Wege führen nach Rom*"/ "*All roads lead to Rome*"/ "*Banyak jalan menuju Roma*" and "*Hunde bellen, die Karawane zieht weiter*"/ "*Dogs bark, but the caravan goes on*"/ "*Anjing menggonggong, kafilah berlalu*".

In addition, Phrasemes whose phraseological meaning is easy to guess can be chosen, for example "*im Geld schwimmen*" (*mandi uang*) and *jemandem den Weg ebnen* (*memuluskan/meratakan jalan bagi seseorang*).

Finding and modifying the appropriate examples of using the German phraseological expressions

To give an idea of how to use a phraseme, authentic examples are shown in the dictionary. Duden 11 consistently provides authentic examples of the use of phrasemes so that it makes it easier for readers to see the structure of the phrasemes and to understand the meaning of phrasemes better. However, many authentic examples are difficult to understand without extensive knowledge. Another problem, there are authentic examples that are quite long and the structure is difficult for elementary level learners. Therefore, the example made by lexicographers can be a solution because it can be made simpler and shorter. They don't need to create new examples, but can modify existing authentic examples to make them simpler. Subjects, objects and sentence structures can be changed. However, here we need sources who can judge whether the examples are acceptable and natural.

According to Lemnitzer and Zinsmeister (2015), corpora are an important source for usage examples. This possibility is more important, especially when it comes to creating a learner's dictionary. But, there are limits in using corpora in a corpus linguistic study. Corpus size is not always satisfactory, especially when it comes to identifying idiomatic phrasemes. The use of phrasemes, especially idiomatic verbal phrasemes, is very low in various languages (Colson 2007).

For which studying the frequency of phrasemes a large corpus is needed, the World Wide Web is viewed as a practical solution. Colson (2007) explains that the best search engine has access to corpora with more than three billion websites in many languages. With Google (http://www.google.com) you get a result of 34,000 examples for the search term "spilll the beans". A more valid result can be achieved with a very large corpus than with a corpus that contains far fewer words (Colson 2007).

Finding equivalences for the phraseological expressions in Indonesian language

There are numerous collections of proverbs or proverb dictionaries for laypeople on the Indonesian market. As proverbs are taught and learned in schools, many proverbs dictionaries are brought to market. So that collections of proverbs can be sold better, they are often not only sold as collections of proverbs, but also combined, for example, with the rules of Indonesian orthography or other learning objects.

Indonesian dictionaries that only list proverbs or idioms without additional subjects such as *Kamus Ungkapan Bahasa Indonesia* (1997) by Mahayana, Nuradji and Totok Suhardiyanto, *Kamus Ungkapan Bahasa Indonesia* (2002) by Chaer, *Kamus Idiom Bahasa Indonesia* (1993) by Chaer, *Kamus Lengkap Peribahasa Indonesia*. Untuk SD, SLTP, SMU dan Umum (2005) by Abdullah, were getting less and less. The years of publication show that there is no pure Indonesian phraseological dictionary has been published in recent years. This indicates that dictionaries of Indonesian proverbs or phrasemes are difficult to sell in the Indonesian market.

Due to the lack of good and actual Indonesian phraseological dictionaries, it is difficult to rely on printed dictionaries alone to look up German equivalent phrasemes in Indonesian. The Lexikographers need to do online studies and rely on their knowledge of the language to find the exact equivalent. The positive side is, the lexicographers can find more actual equivalents. This online study requires caution. The online source used for validation should be good. It would be better if the sources used were online mass media.

Regarding the usage of phrasemes, corpus-linguistic studies are helpful to check the frequency of phrasemes (Burger 2015). This method is more effective for phraseology research than studying dictionaries or interviewing natives. Studying dictionaries don't always bring clarity, because lexical materials in dictionaries are frequently only adopted from the old dictionaries, so that some phrasemes are no longer used in the current language. Burger (2015) recommends the use of corpus linguistic methods in phraseology research in order to avoid unchecked assumptions. Thanks to a corpus linguistic study, it can be empirically proven that a phraseme is "generally" in use and not based on "intuition" or the individual convictions of the linguist.

Determining phrasemes equivalents is not easy. Phraseological meaning is the main factor in determining equivalents. The form or component of the phrasemes can also be considered when looking for equivalents. Another thing to consider is the frequency of use of equivalent phrasemes. It may be that these phrasemes have the same meaning and form, but the level of popularity is very low and there are phrasemes that are more commonly known. In this case, it is of course important to prioritize the more commonly used phrasemes.

According to Burger (2015), there is still not a single dictionary that has fully taken into account the knowledge and requirements of phraseology research. The reasons for this are the large amount of time required to create and edit dictionaries. According to Burger, dictionary editors are not very sensitive to current demands from linguists.

Müller / Kunkel-Razum (2007) had previously complained that the adoption of phrasemes in dictionaries and their lexicographical processing there lagged far behind the results of their research from the point of view of phraseology researchers and dictionary critics. There are deficiencies in the insufficient consideration of the phraseology in the introductory texts of the dictionaries, in the form of contradictions between the introductory texts of the dictionaries and the treatment of the phrasemes in the alphabetical part, in the inconsistent treatment of the phrasemes, in the explanations of meaning, in the marking of the lemmata as well as in the lexicographical examples.

Linguists and lexicographers agree that bilingual dictionaries cannot meet all the reference needs of dictionary users with two different languages to the same extent, since even the best bilingual dictionaries cannot provide the same amount of information for the two languages (Lubensky / McShane 2007).

The above explanations explain the difficulties of making a phraseological dictionary that meet the demands of phraseology experts. However, by studying phraseology and lexicography, lexicographers can create a phraseological dictionary that meets the demands from phraseologists. As already discussed, the existence of phraseological restrictions make the factors that must be considered in making a phraseological dictionary getting more. By limiting the number of phrasemes taken as lemmata and the use of corpus linguistic methods, lexikographers can focus more on making a phraseological dictionary that is able to meet the demands from experts and dictionary users.

5. References

Abdullah, M. K. (2005): *Kamus Lengkap Peribahasa Indonesia. Untuk: SD, SLTP, SMU dan Umum*. Jakarta: Pustaka Sandro Jaya.

Burger, Harald (2015): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 5., neu bearbeitete Auflage. Berlin: Erich Schmidt Verlag.

Chaer, Abdul (1993): *Kamus Idiom Bahasa Indonesia*. 3. Auflage. Flores: Nusa Indah.

Chaer, Abdul (2002): *Kamus Ungkapan Bahasa Indonesia*. 1. Auflage. Jakarta: Rineka Cipta.

Colson, Jean-Pierre (2007): The World Wide Web as a corpus for set phrasemes. In: Burger, Harald/ Dobrovol'skij, Dmitrij/Kühn, Peter/Norrick, Neal R. (Ed.), *Phraseologie: ein internationales Handbuch der zeitgenössischen Forschung*. Volume 2. 2. Halbband. Berlin/New York: Walter de Gruyter. 1071-1077.

Engelberg, Stefan/Lemnitzer, Lothar (2009): *Lexikographie und Wörterbuchbenutzung*. 4., überarbeitete und erweiterte Auflage. Tübingen: Stauffenburg Verlag.

Ettinger, Stefan/Nunes, Manuela (2006): *Portugiesische Redewendungen. Ein Wörter- und Übungsbuch für Fortgeschrittene*. Hamburg: Helmut Buske Verlag.

Herlina, Ina/Nandzik, Kevin (2015): *Andere Wiese, andere Grashüpfer ... oder Andere Länder, andere Sitten. Indonesisch-deutsche Sprichwörter und Redensarten im Vergleich*. Berlin: Regiospectra.

Koller, Werner. 2007. Probleme der Übersetzung von Phrasemen. In: Burger, Harald/Dobrovol'skij, Dmitrij /Kühn, Peter/Norrick, Neal R. (Hg.), *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. Volume 2. 1. Halbband. Walter de Gruyter. Berlin, New York. S. 605-613.

Korhonen, Jarmo (2007): Probleme der kontrastiven Phraseologie. In: Burger, Harald/Dobrovol'skij, Dmitrij/Kühn, Peter/Norrick, Neal R. (Ed.), *Phraseologie: ein internationales Handbuch der zeitgenössischen Forschung*. Volume 2. 1. Halbband. Berlin/New York: Walter de Gruyter. 574-

589.

Kreuder, Hans Dieter (2003): *Metasprachliche Lexikographie. Untersuchungen zur Kodifizierung der linguistischen Terminologie*. Lexikographica Series Maior. Tübingen: Max Niemeyer Verlag.

Lemnitzer, Lothar/Zinsmeister, Heike (2015): *Korpuslinguistik. Eine Einführung*. 3., überarbeitete und erweiterte Auflage. Tübingen: Narr Francke Attempo Verlag.

Lubensky, Sophia/Mc Shane, Marjorie (2007): Bilingual phraseological dictionaries. In : Burger, Harald/ Dobrovol'skij, Dmitrij /Kühn, Peter/Norrick, Neal R. (Ed.), *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. 2. Halbband. Berlin/New York: Walter de Gruyter. 919-928.

Mahayana, Maman S./Nuradji/Suhardiyanto, Totok (1997): *Kamus Ungkapan Bahasa Indonesia*. Jakarta: Grasindo.

Müller, Peter O./Kunkel-Razum, Kathrin (2007): Phraseographie des Deutschen. In: Burger, Harald/ Dobrovol'skij, Dmitrij/Kühn, Peter/Norrick, Neal R. (Ed.), *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*. 2. Halbband. Berlin/New York: Walter de Gruyter. 939-949.

Römer, Christine/Matzke, Brigitte (2005): *Lexikologie des Deutschen. Eine Einführung*. 2., aktualisierte und ergänzte Auflage. Tübingen: Gunter Narr Verlag.

Schaeder, Burkhard (1987): *Germanistische Lexikographie*. Lexikographica Series Maior. Tübingen: Max Niemeyer Verlag.

Scholze-Stubenrecht, Werner/Haller-Wolf, Angelika (2013): (Duden 11) *Duden Redewendungen. Wörterbuch der deutschen Idiomatik*. 4., neu bearbeitete und aktualisierte Auflage. Band 11. Berlin: Dudenverlag.

DIFFICULTIES IN DOCUMENTING ROOT FORMS OF VOCABULARY FROM ORAL LANGUAGES

Rajiv Ranjan Mahto, S. Mohammad Fayez
Maulana Azad National Urdu University Satellite Campus, Lucknow, India
rajivrmahto@gmail.com; smfayez@gmail.com

Abstract

Every language has at least two major lexical categories - noun and verb. The tribal (indigenous) languages of the Jharkhand state in India present a unique situation w.r.t. verbs in particular. Of the 32 languages from the state only one, Santhali, has been included in the Scheduled list of languages in India. Majority of the remaining languages do not have a script yet and, therefore, remain oral languages. The verbs in these languages show a distinct characteristic. Majority of them in languages like Ho, Khadiya, Mundari, and Santhali etc. are conjugated with a light verb in the non-past and past tenses, the only two tenses available in these languages (Ramswamy, 2007) (Hoffman & Emelen, 1930) , and therefore, are available in the V1 form. However, in some other languages like Kudmali, it is difficult to trace the root form of the verbs as the present form of the same too is a derivational morphological entity. As a doctoral research* on the multilingualism prevailing in the region this researcher hypothesises that a multilingual dictionary will be pivotal in developing the cognition of the IMT children (Skutnabb-Kangas, 2014) from this region and will be helpful in English language learning. The present research paper accounts this researcher's personal experience and highlights the challenges one encounters while documenting the verbs onwards to developing a bi/multilingual dictionary for the non-written languages. The challenges are at two levels at the moment – one, at the level of ascertaining the root form of a lexical item, and two, while documenting it through FLEx due to the same.

Key terms: Tribal (indigenous) languages, Verbs, ITM children, Bilingual dictionary and Kudmali.

1.0 Introduction

This lexicographic exercise has been attempted aiming at the middle school (std. VI, VII and VIII) students for English language learning (ELL) in the state of Jharkhand, India. The state of Jharkhand has a high linguistic diversity for as many as 32 languages mark their presence there. The region has very rightly been called the "melting pot" (Abbi,1997) for linguistic convergence as languages from Austroasiatic (AA), Dravidian and Indo-Aryan (IA) groups are found there. Emeneau, too, as far back as 1956, pointed out the presence of these three language groups and acknowledged the linguistic richness of this region. He particularly uses the term Munda for AA while mentioning the vocabulary borrowing between the three language families (Emeneau, 1956). No doubt, linguists from all over the world take interest in the languages as well as the linguistic scenario prevailing here. However, barring Santhali not any other language from this region could become a scheduled language in the official list of government of India. Most of these languages still remain oral languages and therefore continue to pose difficulties for the learners as well as material developers and teachers. One such language, Kudmali, has been selected for doctoral research with an objective to facilitate ELL through an English to Kudmali bilingual dictionary for the Middle School students in rural Jharkhand. This research paper highlights the challenges and difficulties faced during documenting the lexical items in Kudmali language for the making of a unidirectional bilingual L2 - L1 dictionary. Of many, the difficulties arising in documenting the lemma forms of the verbs, modals and lexical relations have been majorly highlighted in this paper.

1.1 The language, Kudmali

Kudmali (pronounced /kʊɽmɑːliː/ and spelt Kurmali in some literature) is a major language in the Indian state of Jharkhand. It is spoken by the Kudmi (/kʊɽmiː/) community across the Chhotanagpur Plateau region which is largely the present-day Jharkhand, a state which is known as Ruhr of India for its richness in mines and minerals. The Kudmi tribe also has presence in the neighbouring states of Odisha and West Bengal, and therefore, a fraction of speakers of the language are there too. Kudmali being scattered to different regions and coming in contact with several dominant languages is subject to differences in phonology, morphology and lexicon. This language is yet to have a script, and thus, transfers from one generation to another orally. However, applying the UNESCO's criteria (UNESCO, 2003) for language endangerment like intergenerational transfer etc., it is certainly in the endangered category. In terms of classification, it has been classified as an Indo-Aryan language. This classification is unjust, and politically charged. We shall leave the political part as it is beyond the concern of this research and concentrate on the language. The speakers of this language, i.e., Kudmis, are a tribe of Dravidian stock (Grierson, 1927) and (Risley, 1891). This fact asserts that they are inhabitants in the region from the pre-Aryan period. Therefore, it shall be linguistically wrong to demarcate a pre-Aryan linguistic community as the speakers of an Aryan language. Like the other indigenous communities from the region this tribe too has its own linguistic identity and denial of which has led to its identity crisis.

1.1.2 Lack of standardization

Documenting the lexical items from this language has been a great challenge. The language has over a period of time incorporated so much from the dominant languages into itself that many of the actual terms have either been lost or undergone modification. In addition, a major lexical variation is noted for the same signified object among the Singhbhum[1] and Dhalbhum[2] varieties of Kudmali. Therefore, Haugen's (1966) four aspects of language development are very crucial in this context. He pointed out - 1) selection of norm, 2) codification of form, 3) elaboration of function and 4) acceptance by the community as the four important steps for a vernacular to develop as standard (Haugen, 1966). Utmost care was taken in selection of the norm for the lemma and acceptance of the same by the community. Since the trends of migration has been from and into the region as the industrialization took place, the scope for linguistic convergence became wide. And with time, it imbibed lexical items from not one but many languages. Therefore, the present generation of the speech community is not the ideal resource for data collection/ verification (selection of norm). This researcher, thus, depended upon the generation which is grandparents now for documenting the lexicon of the language. This is discussed more in the methodology section.

Lack of standardisation of a language is a big deterrent for lexicographers. In such cases, the same signified object ends up with multiple signifiers. Non-standardisation of the language contributes in complicating the matter for the lexicographer as an object may have another term in a different region for the people of same community. For example, *ðaːmɽaː* in Dhalbhoom[1] is used for 'bull' whereas in Singhbhum[2], it means a 'male calf' of a bovine. Similarly, for 'female calf' it has an inflectional derivation, i.e. *ðaːmɽiː*. Such cognates add to the complicacies for a lexicographer as determining the etymology too is a hard nut to be cracked for an unwritten language. Such variation is noted at the level of pronominals too, as illustrated in table 1 below.

Table 1: The pronominal variation within two varieties of Kudmali.

Pronominals	Singhbhum variety	Explanation	Dhalbhoom variety
I	*mõy*	1SG-NOM	*haːmɪ*
You	*t̪õy*	2SG-NOM	*t̪ʊmiː/t̪õ*
S/he	*õy/uː*	3SG-NOM	*uː*
We	*haːmraː*	1PL-NOM	*hʌmʌn*
You [PL]	*t̪ohoraː*	2PL-NOM	*t̪ohniː*
They	*okʰraː*	3PL-NOM	*okʰniː*

Documenting such dual existence of pronominals within a language adds to lexicographer's woes. From a pedagogical point of view, it puzzles the learner initially and requires him/her to acclimatize with the variety of his/her own. However, these do signal the regional identity of the utterer.

2. Sources the study is based on

The Textbooks of English for class VI, VII, and VIII prescribed by Jharkhand Council of Educational Research and Training (JCERT) were used as the target corpus to be translated. The word bank thence obtained was segregated into POS categories and simple and derived forms. Caluwe & Santen's (2003) highlight the distinction between simple and derived form of words as the relationship between form and content is arbitrary and unpredictable in principle in simple words, whereas, in derived words it is highly predictable. They also emphasise that due to the unpredictable nature of the relation between form and content, reducing information on simple words poses considerable problems at the macro and micro-levels (Caluwe & Santen, 2003). In addition, Geeraerts'(2003) article on 'meaning and definition' has been a great explainer in understanding intentional and extensional definitions of words. Since this work is essentially a 'reception-oriented dictionary' (Hannay, 2003) targeted for ELL, it assumes his revelation, "In order to fully understand why and how any one dictionary differs from another, one has to view the dictionary as essentially a translation- related problem-solving tool for users… ." (Hannay, 2003), as a matter of principle.

3. Methodology

As the target users for this L2 - L1 unidirectional dictionary were the middle school students, it was sensible to select the word bank from their textbook. Lexical items from the L1 for the vocabulary encountered in the textbook/s were elucidated in the classroom by the learners using the language of wider communication (LWC) in that region as the first step. Post segregation into POS of the word bank thus obtained was verified at the community level with elderly speakers through the means of questionnaires and interviews. Such word bank was tested in classroom situations among the students at another school in the same sub-divisional area. It was during these sessions it came to the fore that the stem form of a verb was not being successfully transferred. It was then that the methods of field linguistics were employed and questionnaires involving sentences targeting on extracting noun forms were used (Abbi, 2001). Such data collection helped in zeroing down to the lemma forms of the verbs and modals. The lexical items thus finalised were then archived for documentation in the field language explorer (FLEx) software made available by SIL international. The challenge, however, continued to persist at the pedagogical level as the lemma thus reached at does not figure significantly in their repertoire.

4. Discussion

With no standard variety of a language, the lexicographer's first worry is which variety to be trusted and documented. In addition to identifying the lexical equivalents in the target language, s/he requires to get into a comparison exercise in order to note the similar or dissimilar lexical items between the varieties of the language. The lexical items from the varieties figuring in the intersection within a semantic domain is indeed a sigh of relief, those out of it force the lexicographer to plan the explanations differently for such entries.

4.1 The pronominals and the pre/post-positions

At the onset, the documentation exercise for this research was begun with the pronominals and the pre/post-positions in Kudmali. As illustrated in table 1, the pronominals do show variation amongst the varieties of Kudmali that exist. It is a difficult situation to be in for a lexicographer. The multiple signifiers within a language for one lexical entry, as is the case of the pronominals here, due to the absent of the written form and lack of standardization, puts the lexicographer in a conundrum. Pedagogically, this calls for lexical priming, however, the grounds for the same remain undecided until a variety is agreed upon as the standard. Contrastingly, the post- positions of Kudmali exhibit a different character.

Majority of the post-positions are not separately marked in Kudmali. Rather, they are part of the NP as inflectional suffixes which holds true for its varieties too. While documenting the prepositions this researcher had the revelation that this is a language which does not distinguish between the prepositions ‘in’, ‘on’ and ‘at’. Rather, these three are mapped to an inflectional suffix, i.e., a nasal vowel /*ẽ*/ for the purpose of locative. E.g.,

(1)	*Mõy*	*gʰaːr-ẽ*	*aːhõ*
	1SG.NOM	home-LOC	COP-1SG.PRS

I am at home.

(2)	*poʈ̪ʰiː*	*ʈa*	*ʈebiːl-ẽ*	*aːheɪk*
	Book-3SG.NOM	ART(M)	table-LOC	COP-PRS.3SG

book is on the table

(3)	*ðʰaːn*	*kʰeʈ̪-ẽ*	*aːheɪk*
	Paddy-3SG,NOM	farm-LOC	COP-3SG.PRS

Paddy is in the farm

In (1), (2) and (3), it can be seen that unlike the nominative pronominals, multiple equivalents for one entry as seen in table 1, this language uses a single locative, i.e., /*ẽ*/ for the three prepositions (in, on and at) in English. Similarly, one post-position *dʒãːɹ* is used for both ‘upto’ and ‘until’. However, semantically they differ as it is clear from the examples below:

(4)	*bʰõðo-ĩː*	*aːʈʰ*	*dʒã̃ːɹ*	*poɽho-laːhe*
	3SG.NOUN-ERG	class VIII	upto	study-COP-PST

Bhondo has studied upto class VIII.

(5)	*bʰõðo*	*baːdʒaːr-ẽ*	*pãːtʃ-baːɪdʒ*	*dʒãːɹ*	*roh-eiː*
	3SG.NOUN	market-LOC	5pm -time	until	remain-PRS.INFL.3SG

Bhondo remains in the market until 5pm.

(6)	*bersaː*	*hele*	*dʒãːɹ*	*haːr*	*laːg-ʈ̪ek*	
	Rain-3SG.NOUN		COP	COND	plough	tilling-FUT

Ploughing/tilling will be possible if it rains.

In (4), the post-position *dʒãːɹ* is used to mean ‘end of the process’ and thus, clarifies that the subject in the utterance discontinued education after class VIII. Example (5) which gives the information about a regular activity has the post-position *dʒãːɹ* to indicate the deadline. Although these two post-positional uses of *dʒãːɹ* give a sense of ‘limiting’ about the activity, the third use in (6) is that of a conditional, and hence, syntactically, functions as a connector between the cause and effect.

It can also to be noted that the meaning ‘class VIII’ in example (4) was derived by mere mention of the numerical eight from the semantic domain of study - used as verb here. This language practically does not use ordinals. Therefore, the translatability of the suffixes like -st, -nd, -rd and -th marking ordinals in

English proves difficult pedagogically. However, it does have the term for ‘frequency’, i.e., *t̪ʰor* which is mapped to ‘times’ in English. Like, *pãːtʃ t̪ʰor* means ‘five times’ in English.

As a third category in the POS, the word *dʒãːɪ* also functions as a command for the verb ‘go’. It is used as a command sentence for the 2SG very much like ‘go’ is used in English. The subject like in English remains silent.

(7) *dʒãːɪ*
Go-2SG.COM
(You) go.

Thus, the word *dʒãːɪ* in Kudmali, enjoys three functional roles - as a postpositional (as shown in (4) and (5)), as a conditional and hence as a connector (as shown in (6)) and as a verb (as shown in (7)). So, it has to be documented in three separate categories.

Another preposition of place ‘to’ is again a complicated item for the English language learners with Kudmali as their L1. A sentence/utterance using ‘to’ with a destination, like, ‘I am going to the market’ does not carry a pre/post-position in Kudmali for that matter.

(8) *Mõy* *baːdʒaːr* *dʒaː-hõ*
1SG.NOM market go-1SG.PRES
I am going to the market.

Interestingly, in another use as a preposition of place ‘to’ does get mapped to a lexical equivalent. For instance, the routes in airways, railways or roadways do involve use of ‘to’ and ‘from’ with their respective places of commencement and culmination of the journey, and they do get marked as shown below:

(9) *baːs* *ʈa* *rãːtʃiː* *lẽ* *raːuːrkelaː*
Bus-3SG.NOM ART(M) Ranchi from-LOC Rourkela
dʒãːɪ *dʒaːɪ*
to-LOC go-3SG.PRS.
The bus goes from Ranchi to Rourkela.

With this, *dʒãːɪ* has a fourth role in the scheme of the Kudmali language. Keeping Hannay’s (2003) viewpoint about dictionary as “essentially a translation-related problem-solving tool for users”, existence of such lexical items with their wings spread into multiple semantic domains as observed in the case of *dʒãːɪ* is a challenge for any translator as well as a lexicographer. The safest route chosen by the lexicographers in such circumstances is to assign all the relevant POS categories for such entries. Since “the vast majority of bilingual dictionaries are both reception and production oriented” (Ibid) and hence, do double up as “bi-directional” (Ibid) L1 - L2 dictionary as well. The basic level English language learners in the present study indeed found words like *dʒãːɪ* causing errors during L2 production. They are not able to quickly decide for *dʒãːɪ* in a sentence to map it to any one of these - up to, till, until or to; and wrongly find them replaceable by and for each other.

4.2 The derivations through inflections

Caluwe & Santen (2003) propose that dictionary entries on derivations should be marked for their most obvious morphological components and contain information on related words. It indeed becomes a difficult proposition in the case of languages which have less or no use of affixes. Kudmali being one such language, one finds it difficult to zero down the morphological counterparts of the affixes used in English.

For example, to note the meaning of 'unhealthy' one needs to know the equivalent for the un- prefix in the target language. Here, *soaːŋ* is the closest lexical item from the target language that can be mapped to 'healthy'. However, there is no equivalent for 'unhealthy' in the target language. In the case of reception-oriented dictionary, Hannay (2003) holds 'understandability' and the 'usability' of the information provided in it as a must. It was also proposed by him (Ibid) that "If necessary, descriptive information should be added where no L1 equivalents exists". Thus, to transfer the meaning of un-prefix in 'unhealthy' one needs to attach a phrasal expression to the stem *soaːŋ*. The probable equivalents, therefore, are – *dʒaːkor soaːŋ niː* and *dʒon ʈaːĩː soaːŋ niː* which do satisfy both the understandibility and usability.

(10) *dʒaːkor* *soaːŋ* *niː*
RPOSS health COP.NEG
Who does not have health.

(11) *dʒon-ʈaːĩː* *soaːŋ* *niː*
PRO-ART(M)-GEN health COP.NEG.
Which has no health.

However, it becomes an issue for the lexicographer and the learners as 'unhealthy' the adverbial is mapped to nominal or nouns (if one incorporates examples). The learners, thus, end up struggling, "Do the adverbs in English become nouns in Kudmali?".

In the case of lexical borrowing, we do not see such complicacies though. For example, for 'dentist' there is an accepted expression indeed –

(12) *ðãːt̪-ek* *daːkðɔr*
Teeth-GEN. doctor.
Doctor of/for teeth.

Here, *daːkðɔr* is already an accepted borrowing from English ever since the introduction of the allopathy as a mode of treatment in the region and does not have any other lexical equivalent for it in the language. So, a noun in English remained a noun in Kudmali, in an NP. Indeed, the pronunciation witnesses variation as - *daːgðor, daːgʈ̪or , daːkʈor,* and *dokʈor* etc.

4.2.1 Role of inflectional elements

Syntactically, Kudmali is an OV language. Many of the verbs that are transitive in English do not exhibit similar characteristics in Kudmali. Therefore, despite being a nominative- accusative language, passive constructions are minimal in it. For example, 'eat' in 'I eat a guava' can be passivized to 'A guava is eaten by me' for its property of transitivity. However, the Kudmali equivalent for 'I eat a guava',i.e.,

(13) *mõy* *ek-ʈa* *soɪʈ̪ʊmbaː* *kʰãːw*
1SG.NOM one-ART(M) guava eat-PR-1SG
I eat a guava.

despite being in the SOV structure is difficult to be converted into passive. One possible factor for the same could be the absence of an independent equivalent of 'by' conjugating with an agentive in the language. In Kudmali, the marking of agentive and ergative is realized mostly through inflectional nasalization of the nominative. E.g. *kolom* (Pen) – *kolomẽ* (with/by pen), *hõsuːwaː* (sickle) – *hõsuːwaːĩː* (with/by sickle) as in the following sentences.

(14) *mõy* *kolom-ẽ* *lek*^{h}*-õ*
1SG.NOM pen-INSTR write-PR-1SG
I write with a pen.

(15) *mõy* *hõsuːwaː-ĩː* *g*^{h}*ãːs* *kaːtl-õ*
1SG.NOM sickle-INSTR grass cut-PST-1SG
I cut grass by sickle.

The inflectional nasalization here in (14) and (15) marks the agentive. It is, thus, difficult to shift the nasal vowel to the nominative as the 1P SING pronoun does not undergo an inflectional or any other kind of change.

This aspect has to be taken up as another research. Indeed, the issue concerning this research paper is the translatability of such inflectional nasalization. How does this get transferred to the students? What do we provide as the equivalent for 'by' and 'with'? Is it - /*ẽ*/ or /*ĩː*/? This becomes problematic while documenting in FLEx. 'By' and 'with' can be described well with the information and examples discussed here, which will certainly suit the linguists. However, it is unthinkable to make the school students realize such metalinguistic aspects in their use.

4.3 Challenges in lemmatization

In terms of vocabulary development, the learner's aim is to learn the new forms from the target language and their meanings in the source language (Hannay, 2003). This journey is from unknown to known.

> What is unknown is a given L2 item, and the user's main problem is usually that she does not fully understand what the item means in the given context and may wish to translate the item into her own language. (Hannay, 2003)

In the journey from unknown to known, verb plays a very crucial role. The learner has a great expectation from the dictionary when s/he consults it to decipher a clause or sentence from the L2. The lemma form of the verb is of paramount significance in the process from unknown to known as it gets into inflectional adjustments as per the TAM (Tense, Aspect and Mood) and the nominatives. However, lemmatization of the verbs for languages like Kudmali is not a single channel process as it augurs for re/investigations at multiple levels. In Kudmali, each nominative, irrespective of noun or pronoun, has an inflected verb as per TAM agreement.

Table 2: Declension of the verb 'say' in habitual present with all the pronominals.

Sl. no.	English	Kudmali	Lemma - suffix break
1.	I say.	*mõy kohõ*	*koh – õ*
2.	We say.	*haːmraː kohiːyo*	*koh - iːyo*
3.	You say. [SG]	*t̪õy kohiːs*	*koh - iːs*
4.	You say. [PL]	*t̪ohraː kohiːya*	*koh - iːy*ɑ
5.	You say. [HON.]	*t̪ohraː kohiːya*	*koh - iːy*ɑ
6.	He says.	*õy koheiː*	*koh - eiː*
7.	She says.	*õy koheiː*	*koh - eiː*
8.	They say.	*ok*^{h}*raː kohot̪*	*koh - ot̪*

Following the declension route of the verb in the above table, one may note *koh* as the stem for 'say' in Kudmali. At the same time it can also be noticed that none of the pronominals conjugate with this lemma form of the verb.

In the examples in table 2, it can also be noted that verb stem has /*k*/ and /*h*/ consonant sounds in the language connected by a vowel sound. It is at the coda of the verb that we see inflectional changes for aspectual concerns. In Hindi, relatively a far more well-placed language of India, the stem is *kʌh* and the lemma is *kʌhnaː* for 'say' as generally given in the bilingual English to Hindi unidirectional as well as bidirectional dictionaries. It can be determined by the examples below as well:

(16) *kʌhnaː* *meraː* *fʌrz* *hɛ*
Say-GER my-1SG.POSS duty COP
To say is my duty.

(17) *mʊdʒʰe* *kʊtʃʰ* *kehnaː* *hɛ*
I-ACC PRO.INDF say COP
I have to say something.

As a linguist, if one looks through the lenses of declension, one returns with the lemmas *koh* in Kudmali and *kʌh* in Hindi for 'say'. However, *koh* as an independent lexical unit does not have a presence in the Kudmali lexicon unlike Hindi (a far more well researched language with flourishing literature) wherein *kʌh* is predominantly available and is one of the most general words. Let this assertion be verified with another commonly used verb like 'wear' for the analysis.

Table 3: Declension of the verb 'wear' in habitual present with the pronominals.

Sl.No.	English	Kudmali	Lemma - suffix break
1.	I wear.	*mõy pĩðʰõ*	*pĩðʰ - õ*
2.	We wear.	*haːmraː pĩðʰiːyo*	*pĩðʰ - iːyo*
3.	You wear. (SING)	*t̪õy pĩðʰẽ/ / tõy pĩðʰiːs*	*pĩðʰ - ẽ & pĩðʰ - iːs*
4.	You wear. (Pl)	*t̪ohraː pĩðʰiːya*	*pĩðʰ - iːya*
5.	You wear. (HON)	*t̪ohraː pĩðʰiːya*	*pĩðʰ - iːya*
6.	He wears.	*õy pĩðʰeiː*	*pĩðʰ - eiː*
7.	She wears.	*õy pĩðʰeiː*	*pĩðʰ - eiː*
8.	They wear.	*okʰraː pĩðʰot̪*	*pĩðʰ - ot*

As can be deduced from the examples from table 3, in Kudmali *pĩðʰ* is the stem for 'wear'. However, like *koh* this does not exist in the lexicon. Instead, *pĩðʰ*ɑːn is a frequently used form as a gerundial use, however, with a modification in meaning, i.e., 'to offer someone to wear'. It is shown in the following examples:

(18) *pĩðʰaːn* *heɪ* *geleɪk*
Wear-GER be PFV
Offering to wear is over.

(19) *pĩðʰaːn* *kore-laːɪ* *jaːɪho*
Wear-GER PURP go-PROG.1PL
Going for offering to wear.

The lemma form achieved through declension process for the verbs - say and wear - do not serve the pedagogical purposes in an ELL set up as they are difficult to be located in the language system of their L1. Unidentifiable stems for the verbs pose a challenge in style labelling. Difficulty in style labelling is also because of non-standardisation of the language. As the lemma of the verb yielded through stemming does not map to a working stem either in the repertoire of the linguistic community or in the literariness of the language, the language documenter or the lexicographer need to resort to other socio-linguistical strategies. Like posing sentences which are gerund primed or those used for statutory forbidding, etc. as highlighted in examples (16) to (19).

Accordingly, to achieve the lemma, sentences with different syntactical positions for the word in the LWC and the state official language were used. The same is illustrated with the verb 'give' in Kudmali, in table 4 below.

Table 4: Lemma for 'give' and 'drink' through syntactic and semantic variation

Sl. no.	English	Kudmali	obtained lemma
1.	To give is not easy.	*ðeɪk ʈa sohoj ni:*	*ðeɪk*
2.	He made me give.	*õy moke ðeya: kora:ʊla:k*	*ðeya:*
3.	Give me money.	*moke ʈa:ka: ðehĩ:*	*ðehĩ*
4.	He neither gives nor takes.	*õy leya: ðeya: ni: koreɪ* & *õy leɪk ja:ko ni: ðeɪk ja:ko ni:*	*ðeya:* & *ðeɪk*
5.	He made me drink.	*õy moke pɪya: kora:ʊla:k*	*pɪya:*
6.	He paid for my drink.	*õy mor pi:yek k^{h}orca: ðela:k*	*pi:yek*

The lemmata reached at for 'give' in table 4 are *ðeɪk, ðeya:* and *ðehĩ:* of which the first two do not conjugate with the pronominals and only *ðehĩ:* is conjugated with *t̪õy* (You) in a command with the silent subject very much like English. To provide this lemma form as the meaning of 'give' is an awkward situation in a classroom scenario as the instructor sounds like giving the command. The lemma for 'drink' too show the similar inflectional derivation.

On the other hand, mapping the inflected form conjugated with a subject, irrespective of noun or a pronoun, is easily mapped to their equivalents in the L2 and appears relatively more fathomable to the learners than the lemma form of the verb. Thus, both the intentional and extensional definitions of words (Geeraerts, 2003) have to be included during the compilation of a bilingual dictionary like this.

4.4 Expressives translated as reflexives

Native speakers of a language operate with a large number of collocations which contribute to idiomaticity and fluency of their expression while foreign learners do not seem to perceive collocations as chunks and often produce them by combining separate words that do not go together in a given language (Laufer, 2011). It is very much reflected in the expressives of Kudmali language. E.g.,

(20)	*ok^{h}ra:*	*ra:t̪a: ra:t̪i:*	*pa:ra:i*	*gela:*
	3PL.NOM	night-EXPR	leave-PFV	COP.PST

They left in the night itself.

The expressives in Kudmali are formed through derivations. As shown in (20), *raːt̪* night undergoes an inflectional derivation to form *raːt̪aː* which as a separate entity neither has any use nor meaning. The derived form thus obtained further undergoes derivation and forms *raːt̪iː* as its expressive pair. Translation of such items is not quite possible to English as ‘night’ as a noun is almost a closed category, barring the -s suffixation to indicate plurality.

5. Results and findings

Following are the findings of this lexicographic exercise:

- Difficulty in documenting due to non-standardisation of the language. It results into existence of cognates with multiple meanings in the same semantic domain across the varieties.
- Difficulty in determining the derived form in the L1. This contributes to non-inclusivity of the meaning in the L1 as shown in section 4.2 and the examples (10) and (11) about how to describe ‘unhealthy’ as a construction?
- The inflectional suffixes to mark ergativity and genetives in Kudmali are not only difficult to document but also hard to be transferred to the learners as these nasal vowels are bound morphemes (shown in 4.2.1).
- Difficulty in determining the lemma of a verb. As shown through tables 2, 3 and 4 in section 4.3 the lemma obtained through linguistic process of declension does not have a presence in the linguistic system of Kudmali.
- The problem in lemmatization is also reflected at the level of pedagogy as the process of declension becomes difficult to transfer to the learners. Therefore, documentation at phrasal level, in conjugation with nominative or pronominal, is recommended.
- The expressives in this language are difficult to find equivalents for in English and are conveyed through different categories as shown in 4.4 through example (20).

As concluding remarks, it can be said that for someone coming with a linguistic system where pre/post-positions have separate markers it is difficult to be accustomed to the inflectional derivations for the same purpose. A Hindi L1 speaker in a multilingual set-up will find it difficult to transfer the notion of pre/post-positional items to the speakers of tribal languages like Kudmali. Particularly, due to the morpho-syntactic typicalities of these languages as well as due to the lack of maturity to identify and understand these notions by the school going learners. Keeping the pedagogical objective in mind one needs to side with Hannay’s (2003) proposition, “Usability means that once the user has found the L2 item which she wants to use, she must be given information on how to use it”.

Notes:

1. Dhalbhoom used to be a geographical area majorly inhibited by the Kudmis, Santhals etc. till the pre-independence time. It comprises of Puruliya in West Bengal and, Dhanbad and Bokaro districts of present-day Jharkhand. It is largely a colliery area, and hence, witnessed great trends of migration for coal mining and development of market with time.
2. Singhbhum is a mineral rich region in Jharkhand. Also called Ho-land (pronounced homophonically with Holland) due to the dense presence of the Ho tribe. This region is heavily rich in minerals. Ores of iron, copper, mica etc are mined here. This area has been contributing a lot ever since industrialization took place in India. Apart from Ho, Kol and Kudmi tribes form the major chunk of the population here. Presently, it is bifurcated into Singhbhum East and Singhbhum West.

* This research work is part of the doctoral research done by the first author.

References:

Abbi, A. (1997). *Languages of Tribal and Indigenous Peoples of India: The Ethnic Space.* New Delhi: Motilal Benarsidas.

Abbi, A. (2001). A Manual for Linguistic Fieldwork and Structures of Indian Languages. Muenchen: Lincom Europa.

Caluwe, J. d., & Santen, A. v. (2003). Phonological, morphological and syntactic specifications in monolingula dictionaries. In P. v. Sterkenburg (Ed.), *A Practical Guide to Lexicography.* Amsterdam/ Philadelphia: John Benjamins.

Devy, G. N., Gupta, R., & Singh, P. K. (Eds.). (2018). *Peoples' Linguistic Survey of India: The Languages of Jharkhand* (Vols. 13, Part 2). NEW DELHI: Orient Black Swan.

Emeneau, M. B. (1956). India as a Linguistic Area. *Language, 32*(1), 3-16. Retrieved Jan. 03, 2017, from http://jstor.org/stable/410649

Gass, S. (1999). DISCUSSION: Incidental Vocabulary Learning. *Studies in Second Language Acquisition, 21*(2), 319-333. Retrieved from www.jstor.org/stable/44486442

Geeraerts, D. (2003). Meaning and definition. In P. v. Sterkenburg (Ed.), *A Practical Guide to Lexicography.* Amsterdam/Philadelphia: John Benjamins.

Grierson, G. A. (1927). *Linguistic Survey of India* (Vol. V (Part II)). New Delhi: Reprinted by Motilal Benarsidas 1967.

Hannay, M. (2003). Types of Bilingual Dictionaries. In P. v. Sterkenburg (Ed.), *A Practical Guide to Lexicography.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Haugen, E. (1966). Dialect, Language, Nation. *American Anthropologist, 68*(4), 922-935. Retrieved Jan. 04, 2017, from http://www.jstor.org/stable/670407

Hoffman, J., & Emelen, R. V. (1930). *Encyclopaedia Mundarica.* Patna: India: Government Printing Press.

Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. *International Journal of Lexicography, 24*(1), 29-49. Retrieved from https://doi.org/10.1093/ijl/ecq039

Ramswamy, N. (2007). *Ho Grammar.* Mysore: CIIL (Central Institute of Indian Languages).

Risley, H. H. (1891). *Tribes and Castes of Bengal* (Vol. I & II). Calcutta: Bengal Secretariat Press. Retrieved from https://indianculture.gov.in/rarebooks/tribes-and-castes-bengal-vol-i

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, 329-363. doi:10.1177/1362168808089921

SIL. (n.d.). *www.glossary.sil.org*. (SIL International) Retrieved January 14, 2021, from Glossary of Linguistic Terms: https://glossary.sil.org/term/lexical-relation-set-pairs-structure

Skutnabb-Kangas, T. (2014). The Role of Mother Tongues in the Education of Indigenous, Tribal, Minority and Minoritized Children: What can be done to Avoid Crimes against Humanity? In P. W. Orelus (Ed.), *Affirming Language Diversity in Schools and Society. Beyond Linguistic Apartheid* (pp. 215 - 249). Routledge. doi:10.13140/2.1.3350.9764 *The Leipzig Glossing Rules:Conventions for Interlinear morpheme-by-morpheme gloss.* (2015, May 31). Retrieved Feb. 26, 2021, from www.eva.mpg.de: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

UNESCO. (2003). Language Vitality and Endangerment. *International Expert Meeting on the UNESCO Programme Safeguarding of Endangered Languages, Paris, 2003.* Paris: UNESCO. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000183699

Zhang, S., Xu, H., & Zhang, X. (2020, May 27). The effects of dictionary use on second language vocabulary acquisition: A meta-analysis. *International Journal of Lexicography, 34*(1), 1-38. doi:10.1093/ijl/ecaa010

LEXICOGRAPHY AND DOCUMENTATION IN A MULTILINGUAL ENVIRONMENT

Rufus H. Gouws
Stellenbosch University, South Africa
rhg@sun.ac.za

1 Introduction

The significant documentation role of dictionaries as storehouses of language should never be underestimated. Although it is generally accepted that dictionaries document the lexicon or at least certain sections of the lexicon of a given language or language pair this is not the only lexicographic documentation assignment and responsibility. The extent of lexicographic documentation goes beyond the mere reflection of the lexicon, important as this might be. Documentation could also present e.g., typical language use, cultural as well as pragmatic data. In the planning and compilation of their dictionaries lexicographers need to take cognizance of this component of the lexicographic process and the best possible approaches that can suffice the needs of their intended target users. They have to plan the data distribution structure of their dictionaries in such a way that a variety of documentation possibilities can be negotiated, and the information retrieval structure of the dictionaries should allow the target users of to access the documented data in order to retrieve the required information.

2 Background

According to Tom McArthur (1986:19), and with this reference I pay tribute to our friend and colleague who passed away last year, 30 000 years ago Cro-Magnon peoples of Europe began to create drawings on rock surfaces. Caverns with paintings were “public centres of reference” and “a place of reference to which people could come for specific kinds of experience”. This was possible because of the documentation found in these places of reference. Dictionaries are reference sources that reflect much of a language, its people and its cultures. As instruments of documentation they can also be regarded as public centres of reference. This is a role that dictionaries have played for many years. McArthur (1986:24-25) refers to around the year 2350 when the Akkadians invaded Mesopotamia. Sumerian, the language of the people of Sumer, eventually died out because of the pressures of Akkadian. However, earlier lexicographic documentation retrieved from debris surviving from the scribal schools presented archaeologists with clay tablets that reflect lists of Sumerian word-forms with their Akkadian equivalents. This documentation did not only enable the invading Akkadians to become more familiar with the language of the invaded territory but much later it also enabled a study of that language even after it became extinct.

Lexicographic documentation is a process that presents data so that current and future members of a speech community and other interested persons can access a variety of data regarding a given language. It is also aimed at benefitting future scholars needing insight into extinct linguistic and other forms.

In lexicography the notion of documentation is often associated with the compilation of comprehensive dictionaries that cover as big a section of the lexicon as possible. As formulated in the *Worterbuch zur Lexikographie und Worterbuchforschung/Dictionary of Lexicography and Dictionary Research* (WLWF) (2017:83) (= Wiegand et al. 2010-2020) documentation lexicography is: “that branch of language lexicography that has its primary assignment in documenting the full lexicon of a language as comprehensively as possible.” The WLWF contrasts documentation lexicography to usage lexicography that is primarily directed at the individual punctual information needs of individuals, especially lay users.

In this paper documentation in lexicography is not restricted to the efforts of comprehensive dictionaries. As with the Sumerian-Akkadian word lists and more concise forms of the gathering of lexical data are also regarded as forms of lexicographic documentation.

The emphasis in this paper will be on some aspects of the role of lexicographic documentation in a multilingual environment. The South African situation will be taken as a case in point.

3 Containers of knowledge

When discussing reference technologies McArthur (1986:19) uses the expression *containers of knowledge*. This expression can also be applied in lexicography with each member of the wide typological spectrum of dictionaries to be regarded as a container of knowledge and as a carrier of documented data. According to Leibniz (1983: 17) the foundations and base of a language are the words. When a dictionary includes and treats these words it fulfils an important documentation assignment that reflects something of the base of a language. This is important in a multilingual and multicultural society where dictionaries have various assignments in reflecting among others both language and culture - to assist mother-tongue speakers as well as non-mother-tongue speakers who need to retrieve the information contained in the specific containers of knowledge.

In the following sections a number of dictionaries from the South African environment will be mentioned and briefly discussed in order to illustrate various aspects of as well as something regarding the nature and extent of lexicographic documentation.

4 The multilingual South Africa

Within the multilingual South Africa the different languages show an uneven and diverse development pattern that has also had a massive impact on the development (and nondevelopment) of lexicographic work.

Prior to 1994 South Africa had only two official national languages, i.e. Afrikaans and English, whereas a number of indigenous languages were acknowledged on regional level. Since 1994 South Africa has 11 official languages with South African Sign Language currently being considered as a twelfth official language. The eleven official languages and their number of speakers are as follows:

Language	**1st language speakers**	**%**
isiZulu	11,6 million	22,7%
isiXhosa	8,1 million	16%
Afrikaans	6,9 million	13,5%
English	4,9 million	9,6%
Sesotho sa Leboa	4,6 million	9,1%
Setswana	4,1 million	8%
Sesotho	3,8 million	7,6%
Xitsonga	2,3 million	4,5%
siSwati	1,3 million	2,5%
Tshivenda	1,2 million	2,4%
isiNdebele	1,1 million	2,1%

Such a multilingual environment can be regarded as a linguistic and lexicographic laboratory that offers lexicographers golden opportunities to experiment and put their dictionaries to the best possible use. In South Africa the value of dictionaries in a multilingual environment had been acknowledged and has led to the establishment of a government-funded national lexicography unit (NLU) for each one of the official languages. The main task of the NLUs is to compile dictionaries and to document the languages of South Africa.

Different lexicographic projects in different South African languages before and since the establishment of the NLUs give evidence of different approaches to the role of lexicographic documentation and its relationship with the genuine purpose of dictionaries.

According to Al-Kasimi (1977:1) "Each culture fosters the development of dictionaries appropriate to its characteristic demands." This also applies to the South African environment and these demands include the assignment of documentation. However, this assignment is executed in different ways and by means of different types of dictionaries, as will be seen in the subsequent sections.

5 Documentation and the genuine purpose of a dictionary

5.1 Some Afrikaans dictionaries

When evaluating the role of any dictionary as an instrument of documentation one also needs to take cognizance of the intended target users and their needs as well as the genuine purpose of that dictionary. This should have an influence on the nature and extent of documentation achieved in any given dictionary. Different types of documentation can be found in different types of dictionaries with different genuine purposes.

It is possible that similar data are documented but aimed at achieving different purposes. This can be seen in the first two Afrikaans dictionaries, i.e. the *Proeve van Kaapsch Taaleigen* (Changuion 1844) and the *Proeve van een Kaapsch-Hollandsch Idioticon met Toelichtingen en Opmerkingen betreffende Land, Volk en Taal* (Mansvelt 1884). These dictionaries were the first to document Afrikaans as an emerging language. Both were compiled by Dutch linguists who worked in South Africa and whose research was directed at differences between Dutch and Afrikaans. In this regard they had different approaches to Afrikaans, cf. Gouws (2005:96), and these approaches motivated their lexicographic work, but did not diminish the impact of their dictionaries as instruments of documentation. The choice of lemmata in both dictionaries tried to reflect something of the differences between Dutch and Afrikaans and although Mansvelt's dictionary had a bigger macrostructural coverage than that of Changuion, there is a substantial overlap in the lemma selection and many similarities in the treatment allocated to these lemmata. Changuion's work was a significant contribution to the documentation of Dutch dialectal variants, but that was not his main aim. Changuion was not impressed by the changes Dutch had undergone in the Cape and in the preface to his dictionary he clearly states that the main purpose of his work was to rid the Dutch, spoken in South Africa, from the "corrupt" words and expressions he encountered here. It was documentation directed at a form of linguistic cleansing. Contrary to Changuion's approach Mansvelt appreciated the changing linguistic forms, realising that the forms used in South Africa were not only dialectal differences but the emergence of a new language. Mansvelt included four categories of words in his dictionary:

- Words coined in South Africa.
- Dutch words that had acquired a new meaning in South Africa.
- 17th century Dutch words that were still frequently used in South Africa but had
- become obsolete or acquired a highly infrequent use in the Netherlands.
- Words representing language borrowing from other languages like Malay.

In this dictionary lexicographic documentation was employed to display the unique character of an emerging language.

This kind of documentation continued and since 1844 and 1884 the lexicography of Afrikaans expanded and numerous dictionaries representing a wide typological range had been completed. Each one of these dictionaries play a specific role in terms of their documentation objective. The most noticeable in this regard is the work done in the planning and compilation of the multivolume comprehensive *Woordeboek van die Afrikaanse Taal* (the WAT) (Dictionary of the Afrikaans language). Work on this ongoing project was started in 1926 and according to the latest plan it should be completed in 2028. The typological nature of this dictionary resulted in it being comprehensive in various ways - in terms of its macrostructural selection, in terms of the data types presented in the treatment of the different lemmata and in terms of the extent of the treatment prevailing in the different search zones of each article. Especially the first two categories play an important role in the documentation assignment of this dictionary - documenting as large a section of the lexicon of Afrikaans as possible and documenting as many linguistic features as possible of each lexical item selected for inclusion.

5.2 English dictionaries

The preface to Pettman's *Africanderisms. A Glossary of South African colloquial words and phrases and of place and other names* (1913) starts with the following paragraph:

> "When, by some strange oversight, the great 'Oxford Dictionary' not only omits to notice such recognized English words as African and Africanism (Milton, 'Of Reformation in England,' Book 1), to say nothing of such well-known South African words as Africander, Africanderism, and Africanderdom, there does appear to be an excuse, if not a reason, for the publication of a Glossary of South African Words and Phrases."

Pettman arrived in South Africa in 1876 for a stay of nearly forty years. On the day of his arrival, he already jotted down a few of the strange words he encountered in South Africa. That was the unintended beginning of the work on his dictionary. In his dictionary he designates Dutch words and idioms in use in South African English as *Africanderisms*.

In this book Pettman managed not only to display a wide selection of lexical items that he could identify as Africanderisms, the documentation of a specific subsection of the lexicon, but in his treatment of these words he presented numerous examples of their actual use. This added value to the documentation of this work because his dictionary displayed both the lexical stock of this component of South African English and the typical use of these specific forms.

To a certain extent the dictionaries of Pettman and Mansvelt have a comparable genuine purpose: documenting a unique form of a language as it is used on another continent. This kind of documentation firstly is of value to linguists and members of the speech community at the time of the compilation of the dictionary, but it gains value for future generations of linguists interested in the development of a language or variety. The 17th century Dutch used in South Africa developed into a fully-fledged language and the dictionaries of Changuion and Mansvelt now present researchers with the foundation of the early development of this language. English in South Africa did not develop into a new language but into a distinct variety, i.e. South African English. Pettman documented the foundation of this variety.

The documentation of both a language and a variety should be regarded as an ongoing process. This is seen in the lexicographic endeavours of the National Lexicography Unit for South African English. Prior to the establishment of the NLUs the Rhodes University in Grahamstown already hosted a dictionary office working on the lexicographic documentation of South African English. The early work resulted in various editions of *A Dictionary of South African English* (Branford and Branford 1991) and later, after the establishment of the NLU, the *Oxford South African Concise Dictionary* was published in 2002. This dictionary is based on the tenth edition of the British *Concise Oxford Dictionary* (Pearsall 1999) and words typical of South African English complement the general English forms. As such this dictionary documents a selection of items from a specific variety but these items are not dominating the macrostructural coverage.

From a lexicographic documentation perspective, the most significant recent publication dealing with South African English is *A Dictionary of South African English on Historical Principles* (DSAEHP) (Silva 1996). This dictionary, as the cover page indicates, reflects "South African words and their origins". These are South African words belonging to South African English. In its documentation of language this dictionary deviates from a kind of prescriptive approach that only presents good language or the language of the higher social classes and endeavours to present "a selection of exemplary linguistic forms on the basis of canonical authors ... that gave direction regarding good language use" (cf. Gouws, Schweickard and Wiegand 2013:2). In dictionaries with such an approach colloquial language, specialised language and vulgar forms of expression often are taboo. This is not the approach in the DSAEHP.

Irrespective of the type of dictionary, lexicographic documentation is a scientific process that must be performed in an objective and impartial way. It may not be impeded by the lexicographer's personal beliefs, bias or purism. This implies that in the planning of a dictionary lexicographers have to make provision for the inclusion of all lexical items falling within the scope of the assignment of the specific dictionary. This includes taboo words and expressions. As a container of knowledge, a dictionary should also assist its target users in retrieving information regarding taboo words and expressions. This is illustrated by the well- known anecdote of an affected lady telling Samuel Johnson that she highly approved of his not having admitted any improper words into his work. His response to her was: "What, then, I suppose, madam, you have been looking for them." People are looking for all kinds of words and the documentation of a language is incomplete if the writing of a dictionary is impeded by a purist bias. In this regard Nomdedeu Rull (2020:42) complains about Spanish dictionaries not including the frequently used expression *Mucha mierda!* (=Much shit!) to wish someone good luck because it contains a taboo word and such a swear word is considered offensive in the Spanish lexicographic tradition. Lexicographic traditions should not influence the ongoing lexicographic practice of documentation in a negative way.

The editors of the DSAEHP consciously strived not only to document the language of "powerful men" but also the daily speech of ordinary people. (DSAEHP: vii). In the preface it is stated that South African English is not the property of only its relatively small number of English-speakers but also of the greater number of people using English as a second or third language. The DSAEHP also documents the language of these speakers.

The value of the documentation in this dictionary is not restricted to its primary target users but also to researchers in the field of the varieties of English. To them this dictionary shows not only unique South African English words and expressions but also everyday words of English that "are used here in senses which are perplexing to English-speakers elsewhere." (DSAEHP: viii).

Documentation in this dictionary has a further value. Due to South Africa being multilingual there is continuous language contact that results in borrowings between all the languages. The other languages do not only borrow from English, but English also borrows from them. The historical approach in this dictionary also enhances the lexicographic treatment of these borrowed forms in South African English. Linguists from the other South African languages consult this dictionary to retrieve information from its historical data regarding borrowings from their languages because in many ways this dictionary gives a more comprehensive reflection of etymology, and, especially, a chronological presentation of documented usage examples than found in dictionaries of the other languages.

In recent years, the NLU for South African English has made the DSAEHP available as a fully-fledged online dictionary. This enhances the documentation possibilities, as will be seen in the next section.

6 Increasing the extent of lexicographic documentation

6.1 Corpora

Both the WAT and the DSAEHP use corpora to obtain lemma candidates as well as examples of typical usage. The aim of these dictionaries to provide wide-ranging assistance to their users can be reached in an even more effective way by means of additional methods of documentation.

In the microstructural treatment of both these dictionaries cotext items, example sentences in both dictionaries and also collocations in the WAT, constitute an important data type by means of which various aspects of the typical use of a given lexical item can be documented. The examples in the DSAEHP are presented in chronological order which strengthens the historical nature of this dictionary. The usage examples are retrieved from corpora and this adds another possibility to lexicographic documentation. Irrespective of the type of dictionary lexicographers can only include a limited number of usage examples in a dictionary. Dictionary users might often need more typical occurrences and uses of a given word or expression. Space problems in printed dictionaries restrict the number of examples sentences to be included, and in online dictionaries where space is not such a problem too many examples could lead to data overload that could impede rapid access to the required data. However, online dictionaries do offer lexicographers another way of giving their users access to the data documented in a corpus. By complementing the traditional data-pushing structure prevailing in printed dictionaries with a data-pulling structure, cf. Gouws (2018), a link in a search zone of a dictionary article could guide the user to the dictionary corpus where additional occurrences of the typical use of a linguistic expression could be found, cf. Gouws (2021). The corpus is not part of the dictionary, and its contents cannot be regarded as lexicographic documentation in the strictest sense of the word. However, a link from a dictionary article to the corpus implies an increase in the extent of documentation to which users obtain access via a dictionary. This is a type of second level lexicographic documentation. Users need to be made aware of the fact that whereas the dictionary contains curated data, the corpus contains non-curated data. This does not diminish the documentation value of access to a corpus.

6.2 Outer features

Being familiar with developments in the field of metalexicography can also increase the extent of lexicographic documentation. This can be done in both a dictionary-internal and dictionary-external way. Wiegand (1995:465) argues that dictionaries are carriers of text types. The data distribution structure of a dictionary determines where, i.e. in which text, specific data are accommodated in that dictionary. Printed dictionaries offer venues in different search positions, cf. Wiegand et al. (2013:63), like the *search field* (the central list), the *search area* (the dictionary article) and the *search zone* (a microstructural slot in an article). In addition, Gouws (2018:228) indicates that the dictionary as a whole with both central list and outer texts should be seen as a search position in its own right, a *search region*. The planning and execution of a lexicographic documentation process should go beyond the central list of a dictionary and utilize the search region with its additional venues to accommodate selected data. The frame structure of a dictionary, cf. Kammerer and Wiegand (1998), creates the possibility for lexicographers to include front, middle and back matter texts in their dictionaries. These texts could play a significant role in the documentation process by accommodating data that do not typically fit into the search areas. In fulfilling a cognitive function such texts could accommodate data of a more encyclopaedic nature that could have been regarded as of a non-lexicographic nature when presented in the default dictionary articles.

The third volume of the *Greater Dictionary of* Xhosa, a three-volume trilingual Xhosa dictionary, with Afrikaans and English as the other two languages, has a back matter section that includes 52 addenda as outer texts. The first 20 of these texts deal with various aspects of the grammar of isiXhosa. Addendum 21 is a “Note on the anthropological articles that follow”, and in this text it is stated that the material:

> “is offered for the sake of those users of the Dictionary who may become interested and may desire to know more about the beliefs, legends, folklore, social and religious practices and general mode of living of the amaXhosa.”

The addenda 22-52 document a variety of cultural aspects and present the description in the three languages of the dictionary. Many of these lexical items are included as lemmata in the central list where a default treatment is allocated to them. The data distribution in this dictionary is of such a nature that the treatment in the dictionary articles does not give a detailed account of the cultural values of the words represented by the lemma signs, although the treatment might suffice the needs of the mother-tongue speakers of Xhosa. A cross-reference is included in the article of lemmata representing culture-bound words that guides a user to the relevant text in the back matter section where the specific word is treated as cultural phenomenon in its specific cultural context. From a documentation perspective such an outer texts adds value to the dictionary. This documentation shows that as containers of knowledge dictionaries are not only containers of language but also containers of culture. The documentation in this dictionary is not only in accordance with its cognitive function but in the multilingual and multicultural South African environment it exposes some features of a given culture to members of other speech and cultural communities. The value of this type of documentation should never be underestimated.

Access to documented data does not only have to be on a dictionary-internal level. Online lexicography offers the opportunity to link items or item texts in any search position within a dictionary as search region to dictionary-external sources. In this regard Gouws (2018:234) introduced the notion of a *search domain*, i.e. a dictionary portal. From a given item in a dictionary users can be guided by means of a link to other dictionaries populating the same search domain, i.e. the dictionary portal, as the dictionary from which access to dictionaryexternal data is needed. In addition, Gouws (2021) introduced the notion of a *search universe*. This search position is constituted by sources outside a dictionary and its dictionary portal to which the user can be linked from within the given dictionary. The dictionary corpus, other reference sources and even the internet become accessible dictionary-external containers of documented data. This implies that when lexicographers are assessing the documentation possibilities of their work, they need to be aware of existing dictionary-external sources with documentation that can complement that of their dictionaries by means of linking procedures. Access to lexicographic documentation no longer only prevails in the search field of a dictionary but also in its search region, search domain and search universe.

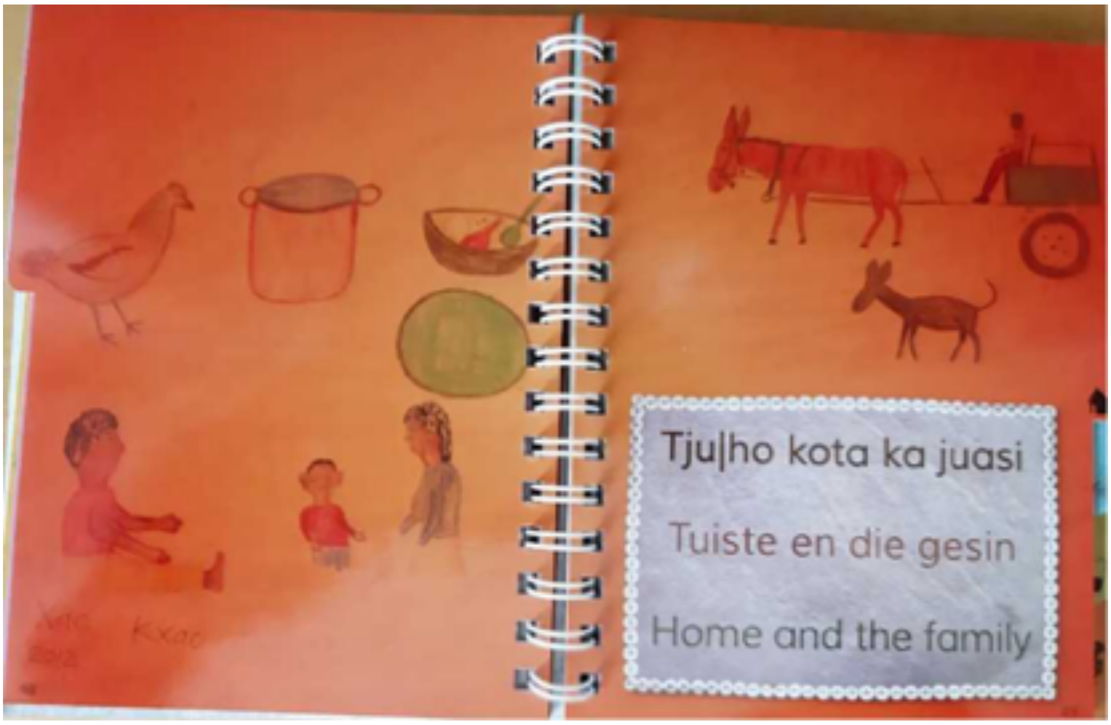
6.3 Endangered languages

Dictionaries can play a vital role in the documentation of endangered languages. This role is enhanced if the documentation includes a focus on culture and even more if it reflects a specific community involvement.

Ju|'hoansi, a Northern Khoesan language spoken in Botswana and Namibia, is an endangered language with only 11 000 speakers left. The *Ju* I' *hoan Tsumkwe Dialect / Prentewoordeboek vir kinders / Children's picture dictionary* (Jones et al. 2014) gives evidence of a diverse approach to its documentation assignment. In the first instance, its lemma coverage documents a part of the lexicon of Ju|'hoansi, but the lemma selection documents yet another aspect of the language. The macrostructure of this dictionary displays a thematic ordering with article stretches representing semantic fields like animals, birds, insects, reptiles and creepy crawlies, home and family, hunt, gather and dance. The macrostructural coverage documents words from these fields. On the back cover of the dictionary a guiding slogan of this lexicographic endeavour is formulated as: "Hold your people, your language and your culture tightly together."

The lemma selection of the *Ju*I'*hoan* dictionary contributes to the documentation of Ju|'hoansi but it does not do it in a random way. Gouws (2006:25) argues that from a lexicographic perspective *cultural data* do not only refer to the world of art, literature, etc., according to Kavanagh (2000:102) the "culture with a capital C", but also traditional beliefs, and the way language reflects the day-to-day life and view of a given speech community, the "culture with a small c". The important role of dictionaries with regard to the documentation of cultural values also follows from the fact that dictionaries are regarded as identityestablishing prestige objects for the conservation of cultural traditions, cf. Gouws, Schweickard and Wiegand (2013:2). Consequently, in their documentation assignment lexicographers need to realise how important even "little culture" is to lexicography. The thematic fields in the *Ju* I' *hoan* dictionary display a significant part of this little culture and the selection of lemmata documents that part of the lexicon that represents the typical and day to day world of the speakers of this language. This is the part of the lexicon the lexicographers want to preserve.

The articles in this dictionary have a homogeneous article structure and contain only a few items. These items are the Ju|’hoansi word, followed by its Afrikaans and English translation equivalents. A pictorial illustration appears to the left of these items. Each thematic field is introduced by a title page, see Figure 1 for the field “Home and family”.



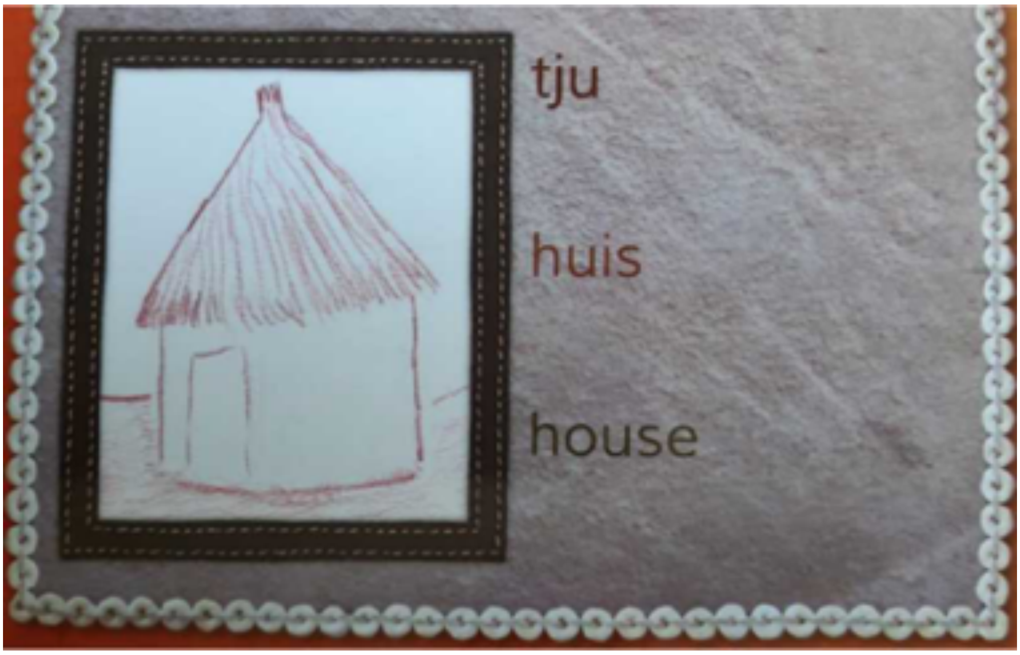
Tju|ho kota ka juasi

Tuiste en die gesin

Home and the family

Figure 1

Within the article stretch of this theme Figure 2 shows the article with *house* as its English equivalent.



tju

huis

house

Figure 2

From a documentation perspective the pictorial illustrations in the articles play an important role in at least two ways. Firstly, in a front matter text the lexicographer says:

> “This dictionary is a collaborative project. This means that many people worked together to draw the pictures, choose the words and record the sound and video clips.”

(By the way, the dictionary also comes with a CD ROM that contains video and audio clips.). This statement confirms the non-random selection of lemmata and that the pictorial illustrations document a certain execution of the lexicographic process. This is the documentation of lexicographic democracy with the target users/speech community directly contributing to the making of the dictionary and thereby making the dictionary not only a joint project but also a joint property. This is not buying into something but rather working into something.

Secondly, the obvious and extreme importance is the way in which these illustrations depict something of the world view of the Jul’hoan speakers. This is not only a picture of a house but this picture documents their understanding of the concept “house” better than a paraphrase of meaning, a mere translation equivalent, a photo from the internet or a drawing by an artist giving his/her own impression of the concept of a house could have done. This picture documents the real-world view of the real speakers of the Ju|’hoansi speech community as to what they regard as a house. This is a significant contribution to the documentation process.

The Ju|’hoan Tsumkwe Dialect / Prentewoordeboek vir kinders / Children’s picture dictionary is a splendid example of the documentation of the lexicon of an endangered language, the world view of the speakers of that language and the eagerness of the ordinary member of the speech community to contribute to the conservation of their language. “Hold your people, your language and your culture tightly together.”

6.4 Dictionary plus

The role of dictionaries in the process of language documentation is not restricted to a traditional form of the lexicographic practice. Innovative forms of documentation come to the fore, e.g. where the lexicographic work is complemented in a single source with other forms of documentation resulting in a product that can be regarded as a dictionary+. This type of documentation can be found across the typological spectrum of dictionaries. I will refer to only two examples of a dictionary+ approach.

6.4.1 Documenting an endangered language

N|uu, one of the few surviving non-Bantu click languages in Southern Africa is one of the most endangered languages on the continent. Two sisters are the only fluent speakers of this language. For the last 14 years one of these sisters and her granddaughter have been engaged in teaching N|uu to descendants of the original speech community. The Centre for African Language Diversity at the University of Cape Town is supporting these efforts by making educational materials available. A most significant contribution in this regard is an illustrated trilingual (N|uu, Afrikaans, English) reader: *Ouma Geelmeid ke kx’u ||xa||xa N|uu/ Ouma Geelmeid gee N|uu*. (= Granny Geelmeid teaches N|uu) (Shah and Brenzinger, 2016). This reader is divided into chapters in which words and expressions from a number of different thematic fields are presented, along with a few illustrations. In these thematic sections a variety of expressions are given in N|uu with translations into Afrikaans and English. In addition to the expressions illustrating the typical use of the language some chapters, e.g. the chapter dealing with animals, also contain words from that semantic field with an illustration for each word.

By giving the expressions the reader adheres to a text production and translation assignment whereas the pictures satisfy a text reception and cognitive function. The lexicographic component is explicitly realised in two glossaries, N|uu-Afrikaans-English and Afrikaans- N|uu-English, presented as the final texts in this carrier of text types. These glossaries are preceded by illustrated charts of the various clicks, consonants and vowels of N|uu.

This reader is not a dictionary in the traditional sense of the word, but it contains lexicographic components complemented by other texts that present lexical, phonetic, orthographic and syntactic documentation of this endangered language.

The reference in Wiegand (2013:285) to printed utility tools with formal properties of lexicographic nature also applies to this dictionary. The principles of language documentation typically found in lexicographic work dominate this publication and the application of established lexicographic principles resulted in an innovative source of language documentation.

6.4.2 Documenting words and expressions dedicated to a specific situation of use

Documentation occurs over a spectrum of dictionary types, including those dictionaries dealing with the full extent of the lexicon of a given language and those dictionaries reflecting a specific subsection of the lexicon, e.g. that of a specialised field. An example of a dictionary+ approach where the reference source with formal properties of a lexicographic nature is directed at a subsection of the lexicon of Afrikaans is Van Sterkenburg (2009): *Moenie mounie/Niet mekkeren. Zuid-Afrikaans met een glimlach* (Don't complain. Afrikaans with a smile). This source was compiled as a guide to Dutch tourists visiting South Africa, especially during the final of the Football World Cup in 2010.

The treatment in this dictionary reflects the typical words and expressions relevant to tourists with regard to their day-to-day communication in South Africa. The lexicographic component of this reference source is embedded in a narrative dealing with various aspects of life in South Africa. Two fictional characters are touring South Africa. In each chapter the plus value is presented by means of a brief introductory to the topic as it is relevant within the South African context. The chapters deal with different themes like the history of South Africa, features of Afrikaans, routine formulae, general vocabulary, daily utility tools, in and around the house, traffic, eating and drinking, animals, sport and some more. The reader of this reference source follows the two characters on their journey and obtains valuable information of a cognitive nature from the narrative which forms the context for a lexicographic treatment of relevant words and expressions. The target readers of this book are Dutch tourists. Dutch and Afrikaans are closely related languages and there are many false friends between these languages. One of the chapters offers a list of words "that do not mean what you think they mean" and a discussion of the meaning of these Afrikaans words focussing on the way in which they differ semantically from their Dutch counterparts or the unique meanings of some opaque complex words. This specific chapters offers a documentation of words and expressions extremely relevant to Dutch mother-tongue speakers encountering Afrikaans.

This dictionary+ publication presents different forms of documentation. The typical lexicographic documentation of a dedicated section of the lexicon prevails but it is put into a historical, pragmatic and usage context with the text of the context equally important as a means of documentation.

Moenie mounie/Niet mekkeren. Zuid-Afrikaans met een glimlach illustrates in no uncertain way a form of combinatory documentation - both lexicographic and non-lexicographic documentation prevailing in a single reference source. This book is compiled for the general public and not for academics or lexicographers. It adheres to the words of Zgusta (1971:16) " .. we must never forget that the lexicographer is doing scientific work, but that he publishes it for users whose pursuits are always more practical, ..." With this publication Van Sterkenburg has managed to implement sound lexicographic and reference principles. This does not only give the book a strong scientific basis but it maintains a user-orientation and sets a new example of data distribution possibilities by introducing a contextualised combination of lexicographic, linguistic and pragmatic data. This is innovative documentation.

7 Conclusion

As containers of knowledge dictionaries have an extremely important documentation role to play. Lexicographers need to be aware of this but need also to be aware of the possibilities to increase the extent of the documentation processes. On a dictionary-internal level not only language but also the culture of the speech community should fall within the scope of documentation. In certain projects the documentation could be enhanced by more active contributions from members of the relevant speech community.

In the lexicographic practice more cognizance should be given to discussions in metalexicography. Realising that there are different search positions to which data can be allocated could increase the documentation possibilities of a dictionary. In this regard it is important to work with the dictionary as a whole, the search region, and not only the central list, the search field. Utilising a frame structure could lead to the inclusion of data in the outer texts that could strengthen the position of a dictionary as a documentation instrument.

The transition to online lexicography has created new possibilities that include improved ways of documentation. Linking a given dictionary to other dictionaries in the same portal widens the scope of information that the user can retrieve. Opting for a search universe and the implementation of data pulling procedures can help the user to reach data documented outside the dictionary. A dictionary no longer is the final destination of a consultation procedure but it can become a transit area from where users can be directed to other destinations that offer a documentation of data not to be found in the single dictionary.

Lexicography remains a remarkable, dynamic and vibrant practice in a stimulating field of study. Documentation, one of the oldest assignments of dictionaries, offers lexicographers exciting challenges and the opportunity to contribute in a dynamic way to giving access to knowledge. After many years, the words of JR Hulbert (1955) have not lost their validity: “I know of no more enjoyable intellectual activity than working on a dictionary.” To which I would like to add: “and talking about dictionaries.”

8 References

8.8 Dictionaries

Anon. 1902/04. *Patriotwoordeboek/Patriot Dictionary*. Paarl.

Botha, W.F, 1951 - *Woordeboek van die Afrikaanse Taal.* Stellenbosch: Buro van die WAT.

Branford, J. and Branford, W. 1991^{4}. *A Dictionary of South African English.* Cape Town: Oxford University Press.

Changuion, ANE 1844. Proeve van Kaapsch taaleigen. In: Changuion, A.N.E. *De Nederduitsche taal in Zuid-Afrika hersteld.* Rotterdam

Jones, K.L. et al. (eds.). 2014. *Ju* I’ *hoan Tsumkwe Dialect / Prentewoordeboek vir kinders / Children’s picture dictionary*. Pietermaritzburg: University of KwaZulu-Natal Press.

Kavanagh, K. (ed.) 2002. *Oxford South African Concise Dictionary.* Cape Town: Oxford University Press.

Mansvelt, N. 1884. *Proeve van een Kaapsch-Hollandsch idioticon met toelichtingen en opmerkingen betreffende land, volk en taal*. Cape Town.

Pahl, H.W. (ed.) 1989. *The Greater Dictionary of Xhosa, Volume 3.* Alice: University of Fort Hare.

Pearsall, J. (ed.) 1999. Concise Oxford Dictionary. Oxford: Oxford University Press.

Pettman, C.P. 1913. *Africanderisms. A Glossary of South African colloquial words and phrases and of place and other names.* London: Longmans, Green and Co.

Shah, S. & Brenzinger, M. (eds.) 2016. *Ouma Geelmeid ke kx’u ||xa||xa N|uu/Ouma Geelmeid gee N|uu*. Cape Town: CALDi (=Centre for African Language Diversity), University of Cape Town.

Silva, P. (ed.) 1996. *A Dictionaryof South African English on Historical Principl*es. Cape Town: Oxford University Press.

Van Sterkenburg, P.G.J. 2009. *Moenie mounie/Niet mekkeren. Zuid-Afrikaans met een glimlach.* Schiedam: Scriptum

8.2 Other literature

Al-Kasimi, A.M. 1977. *Linguistics and Bilingual Dictionaries.* Leiden: E.J. Brill.

Gouws, R.H. 2005. Lexicography in Africa. Brown, K. (red.) *Encyclopedia of Language & Linguistics*. 2nd Edition. Oxford: Elsevier 2005: 95-101.

Gouws, R.H. 2006. The selection, presentation and treatment of cultural phrases in a multicultural dictionary. *Lexicographica* 22: 24-36.

Gouws, R.H. 2018. Expanding the data distribution structure. *Lexicographica* 34, 225-237.

Gouws, R.H. 2021. Expanding the use of corpora in the lexicographic process of online dictionaries. In Taborek, J. (ed.) *10. Kolloquium zur Lexikographie und Wörterbuchforschung.* Berlyn: De Gruyter.

Gouws, R.H. et al. (eds.) 2013a. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: De Gruyter.

Gouws, R.H., Schweickard,W. & Wiegand, H.E. 2013. Lexicography through the ages: From the early beginnings to the electronic age. Gouws, R.H. et al. (eds.) 2013: 1-24.

Hulbertt J.R. 1955. *Dictionaries: British and American.* London: Andre Deutsch.

Kammerer, M. & Wiegand, H.E. 1998. Über die textuelle Rahmenstruktur von Printwörterbüchern. Präzisierungen und weiterführende Überlegungen. *Lexicographica* 14t 224-238.

Kavanaght K. 2000. Words in cultural context. *Lexikos* 10t 99-118.

Leibnizt G.W. (1983): Unvorgreifliche Gedankent betreffend die Ausübung und Verbesserung der deutschen Sprache. Zwei Aufsätze. Hrsg. v. Pörksent U. Kommentiert v. Pörksent U. u. Schiewet J. (Universal-Bibliothek 7987.) Stuttgart.

McArthurt T. 1986. *Worlds of Reference.* Cambridge: Cambridge University Press.

Nomdedeu Rullt A. 2020. How to select and present cultural data: a challenge to lexicography. *Lexicographica* 36: 39-57.

Wiegandt H.E. 1995. Lexikographische Texte in einsprachigen Lernerwörterbüchern. Kritische Überlegungen anläßlich des Erscheinens von Langenscheidts

Großwörterbuch Deutsch als Fremdsprache. In Popp, H. (Hrsg.) *Deutsch als Fremdsprache. An den Quellen eines Faches.* München: Iudicium, 463-499.

Wiegand, H.E. 2013. Gedruckte Gebrauchsgegenstände mit lexikographischen Formeigenschaften. *Lexicographica* 29, 285-307.

Wiegand, Herbert Ernst, Sandra Beer en Rufus H. Gouws. 2013. Textual Structures in Printed Dic-tionaries. An Overview. Gouws, Rufus H. et al. (eds.). 2013a: 31-73.

Zgusta, L. 1971. *Manual of lexicography.* The Hague: Mouton.

THE WORD WITHOUT CORPUS: DESIGNING THE LEXICOGRAPHY AND SYSTEM ARCHITECTURE OF PROJECT MARAYUM

Samantha Jade Sadural
Department of Linguistics, University of the Philippines Diliman, The Philippines
stsadural@up.edu.ph

Abstract

Project Marayum, a University of the Philippines Diliman and Department of Science and Technology (Philippines) funded project, is designed to be a collaboratively-built, online dictionary platform for Philippine languages. Marayum (http://marayum.ph/) provides preliminary bilingual dictionaries of Philippine-type languages despite lacking substantial corpus collection, which was concluded as the primary source of errors in current Filipino monolingual dictionaries. (Lee, 2010) This paper details the development of the lexicographic process of Project Marayum, from data collection, patterned after language documentation practices, to the generation of additional headwords, definitions, and illustrative sentences from community crowdsourcing, which guided the website's system architecture. Documentary lexicography, a type of dictionary that answers the community's need to preserve information for the future (Svensen, 2009), is employed by using an elicitation process that identifies lexical items by semantic fields through a wordlist. The wordlist, which has an initial 505 words, covered different domains of meaning: material culture, human processes and body features, social structures, cosmology, flora, fauna, food, functional lexical words, and verbs and motion words. The community writers then expanded the wordlist by adding related terms to the initial words, resulting in additional 800-1,000 lexical items used in the seed dictionary. Crowdsourcing is envisioned to be used to expand the seed dictionary into a general-purpose dictionary. Crowdsourcing is a term first introduced in 2006 to signify a process that involves a group of people that contribute towards achieving a goal by distributing the overall workload among the individuals of the group (Howe, 2008). The Marayum website was launched on 16 March 2021. It has four dictionaries with English as its L2: Asi (bno), Hiligaynon (hil), Cebuano (ceb), and Kinaray-a (krj). The following eleven dictionaries are currently being collated using Marayum: Akeanon (akl), Bikol-Buhi'non (ubl), Bikol-Central (bcl), Bikol (bik), Gaddang (gad), Itawis (itv), Ivatan (ivv), Kapampangan (pam), Masbatenyo (msb), and Ilocano (ilo). These dictionaries are being managed by their communities and assigned linguists.

Keywords Collaborative Language Dictionary, Electronic Lexicography, Documentary Lexicography, Content Management System, Philippine languages.

1 Introduction

There are 186 established languages in the Philippines— 34 are in trouble, 11 are dying, and two are extinct (Eberhard, Simons, Fennig (eds), 2021). The loss of vitality occurs when the older generation teaches their language to fewer members of the younger generation (Lewis & Simons, 2010). A dictionary, both a documentation of language use and a study guide, can help transmit and preserve a language for future generations. MTB-MLE stands for Mother Tongue-Based Multilingual Education, a program established through Department Order 16 series of 2012 of the Department of Education (2012). MTB-MLE is implemented in all public schools in the Philippines, specifically in Kindergarten and Grades 1 to 3, as part of the K to 12 Basic Education Program.

Since the student begins their learning with the language that they know best, educators can focus on developing cognitive skills and academic content comprehension from the get-go. Studying in a familiar language also provides a local context to the learning task; the student's beginning education environment can be delivered in the culture and manner already familiar to them. With a strong foundation, students can quickly carry over their knowledge and skills to other languages, particularly Filipino (L2) and then to English (L3), later on in their education. Indeed, "we only learn to read once" (Department of Education, 2013a).

Unfortunately, only 19 languages have been listed for the MTB-MLE. Twelve were defined in the original Department Order (Department of Education, 2012), while an additional seven languages were defined later (Department of Education, 2013b). These are:

- Tagalog (tgl)
- Kapampangan (pam)
- Pangasinense (pag)
- Ilokano (ilo)
- Bikol (bik)
- Ybanag (ibg)
- Sinugbuanong Binisaya (ceb)
- Hiligaynon (hil)
- Waray (wrz)
- Tausug (tsg)
- Maguindanaoan (mdh)
- Maranao (mrw)
- Chavacano (cbk)
- Ivatan (ivv)
- Sambal (xsb)
- Aklanon (akl)
- Kinaray-a (krj)
- Yakan (yak)
- Surigaonon (sgd)

The Department of Education is open to dialog on additional languages to the MTB-MLE program (Department of Education, 2020). Having proper language documentation, such as having a dictionary, can significantly assist these efforts. Unfortunately, the lack of Philippine- type language corpora is a significant problem as a corpus of natural connected discourse is a valuable tool in dictionary-making.

This is where Project Marayum comes in. Project Marayum (http://marayum.ph/), a University of the Philippines Diliman and Department of Science and Technology (Philippines) funded project, provides tools for a language community to create, upload, and maintain their language dictionary without the need for technical training in the lexicographic and information technology fields. This would allow the creation of preliminary or seed bilingual dictionaries of Philippine-type languages despite having a lack of substantial corpus collection, which was concluded as the primary source of errors in current Filipino monolingual dictionaries (Lee, 2010).

By definition, a corpus contains a large amount of naturally occurring language data and is an ideal data source for investigating language and language use. While Marayum follows a bilingual dictionary format with English as its second language, the need for a substantial electronic corpus is indispensable as basis material for the compilation of commercial dictionaries and general-purpose dictionaries (Svensen, 2009). A language corpus is also a vital component in research on various aspects of studying the nature and functions of natural language and its multifaceted applications such as language education, lexicography, and natural language processing (Dita, Roxas, & Inventado, 2009).

However, creating a data bank that will comprise a language corpus will take some time. In creating the *Palito* corpus (Dita, Roxas, & Inventado, 2009), the project was given three months for data gathering and completion. From the initial target of one million words of the top four Philippine languages (Tagalog, Cebuano, Ilocano, and Hiligaynon), as patterned after the International Corpus of English (ICE) design, the time constraint compelled the proponents to reduce the number of words to 250,000 and the spoken aspect of the corpus altogether (Dita, Roxas, & Inventado, 2009). It would be harder to collate data from low-resource Philippine- type languages, especially languages that do not have an official or community-agreed orthography and reference grammar.

Unfortunately, time is a resource that Philippine languages do not have in abundance. According to the Summer Institute of Linguistics, almost 96% of Philippine languages are indigenous. (Eberhard, Simons, Fennig (eds), 2021) However, "the actual number of threatened languages remains as mystery, as no detailed and systematic census has been administered to test the SIL's research in these communities," said Jesus Federico Hernandez, a historical linguist, and professor at the University of the Philippines Department of Linguistics (Reysio- Cruz, 2019). In this century, it is projected that more than half of the world's 6600 languages will become extinct, and most of these will disappear without being adequately recorded (Crystal, 2002).

In response to the apparent lack of corpus, or the unavailability of one, in Philippine-type languages, Marayum used documentary lexicography, a type of dictionary that answers the community's need to preserve information for the future (Svensen, 2009). This method was employed through an elicitation process that identifies lexical items by semantic fields through a wordlist. The wordlist, which has an initial 505 words, covered different domains of meaning: material culture, human processes and body features, social structures, cosmology, flora, fauna, food, functional lexical words, and verbs and motion words. The lexicographic framework used in Marayum following the practical lexicographic approach by Weigand (1998) then became the basis for the project's system architecture.

2 Method

2.1 Community-based Participatory Research

As a community and research project, Marayum uses a Community-based Participatory Research approach that engages the community involved with the issue being studied to improve their well-being (Viswanathan et al., 2004). This type of research has been conducted in various fields and is described in different terms. However, it still follows its basic principle: to engage a given community to participate in every aspect of the research process to produce research relevant to their circumstances (Hacker, 2013). This method also ensures building rapport between the community and the research team concerning the expected outcomes or outputs.

2.2 Practical and Documentary Lexicographic Approach

Lexicography is an activity that consists of observing, collecting, selecting, analyzing, and describing several lexical items (words, word elements, and word combinations) belonging to one or more languages in a dictionary (Svensen, 2009). Marayum originally intended to use the methodologies of practical lexicography (or simply dictionary-making). However, the deploring lack of Philippine-language corpora available necessitated the integration of lexicographic documentary techniques. Documentary lexicography, a type of dictionary that answers the community's need to preserve information for the future (Svensen, 2009), was employed through an elicitation process that identifies lexical items by semantic fields through a wordlist. The wordlist, which has an initial 505 words, covered different domains of meaning: material culture, human processes and body features, social structures, cosmology, flora, fauna, food, functional lexical words, and verbs and motion words. A complete list of 505 words is written in the Annex of this paper.

Through a practical lexicographic approach (Wiegand, 1998, as cited in Schierholz, 2015), Marayum followed and integrated these phases into the documentary lexicography practices:

1. Preparation and planning phase
2. Material collection phase
3. Material processing phase
4. Material evaluation phase
5. Publication preparation phase
6. Publication phase

Different decisions must be taken in each phase, actions must be done, and other methods must be used (Schierholz, 2015). Guided by documentary lexicography, these decisions served as the basis for the project's system architecture. This will be discussed further in the Results section.

2.3 Crowdsourcing

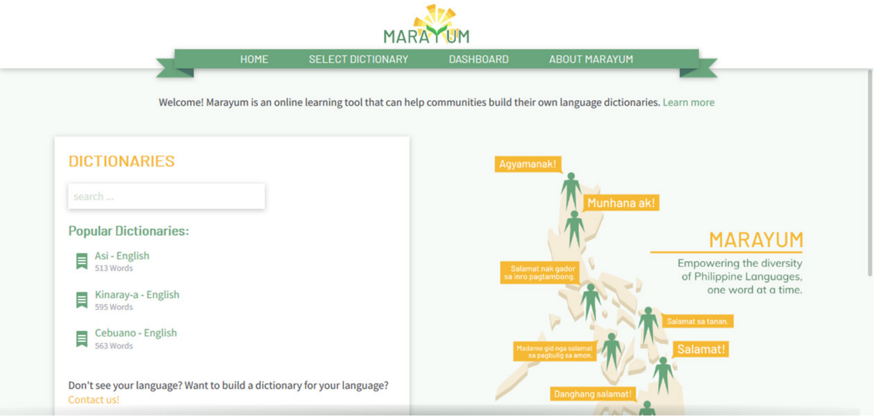
As Marayum is community-based, it will rely on crowdsourcing to generate additional headwords, initial definitions, and sample or illustrative sentences. Crowdsourcing, a term first introduced in 2006, signifies a process involving a group of people (also called a crowd) that contributes towards achieving a goal by distributing the overall workload among the group's individuals (Howe, 2008).

When applied to lexicography, crowdsourcing describes a range of distinct methods for creating or gathering linguistic data (Rundell, 2015). As its language community manages a Marayum dictionary, it functions more as a "wiki model," a crowdsourcing model that, as Lew (2014) asserts, "puts the collective opinion of a group of people above that of a single expert." The model also supports Marayum's vision that the community will make the final decisions about their language's orthography, at the least. A "presiding authority will not make such decisions;" rather, it will be a product of an ongoing collaborative process involving a self- regulating community of contributors (Rundell, 2015).

3 Results

3.1 The Marayum website (http://marayum.ph/)

The Marayum website, which was the actual result of this study, was launched on 16 March 2021. It has four initial dictionaries: Asi-English, Hiligaynon-English, Cebuano-English, and Kinaray-a-English. As of 1 June 2021, the following eleven dictionaries are currently being collated using Marayum: Akeanon-English, Bikol-Buhi'non-English, Bikol-Central-English, Bikol-Rinconada-English, Gaddang-English, Itawis-English, Ivatan-English, Kapampangan- English, Masbatenyo-English, Ilocano-English, and Ivatan-English.



Welcome! Marayum is an online learning tool that can help communities build their own language dictionaries. Learn more

DICTIONARIES

Popular Dictionaries:

- Asi - English 513 Words
- Kinaray-a - English 595 Words
- Cebuano - English 563 Words

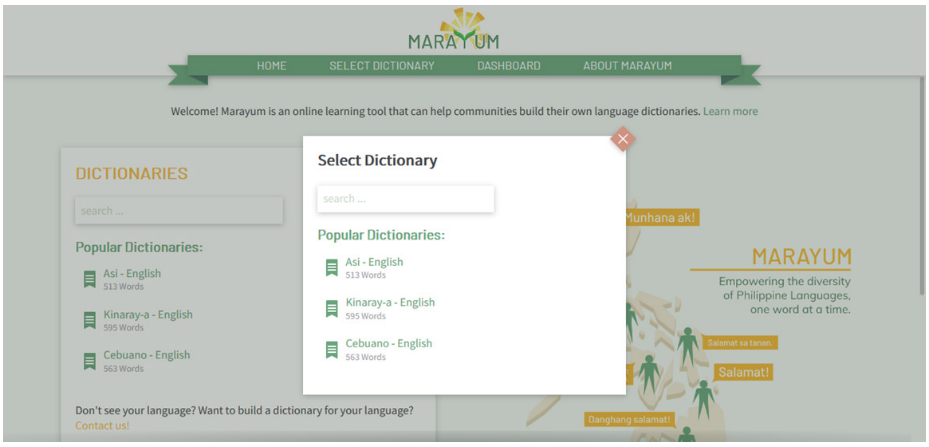
Don't see your language? Want to build a dictionary for your language? Contact us!

MARAYUM

Empowering the diversity of Philippine Languages, one word at a time.

Figure 1: Home Screen

The Home Screen, as shown in Figure 1, is the first page that the user sees when they visit the Marayum website. The screen features a list of top dictionaries in terms of word entries. Featured dictionaries can also be edited in the administrative interface of the website.



Welcome! Marayum is an online learning tool that can help communities build their own language dictionaries. Learn more

DICTIONARIES

search ...

Popular Dictionaries:

- Asi - English 513 Words
- Kinaray-a - English 595 Words
- Cebuano - English 563 Words

Don't see your language? Want to build a dictionary for your language? Contact us!

Select Dictionary

search ...

Popular Dictionaries:

- Asi - English 513 Words
- Kinaray-a - English 595 Words
- Cebuano - English 563 Words

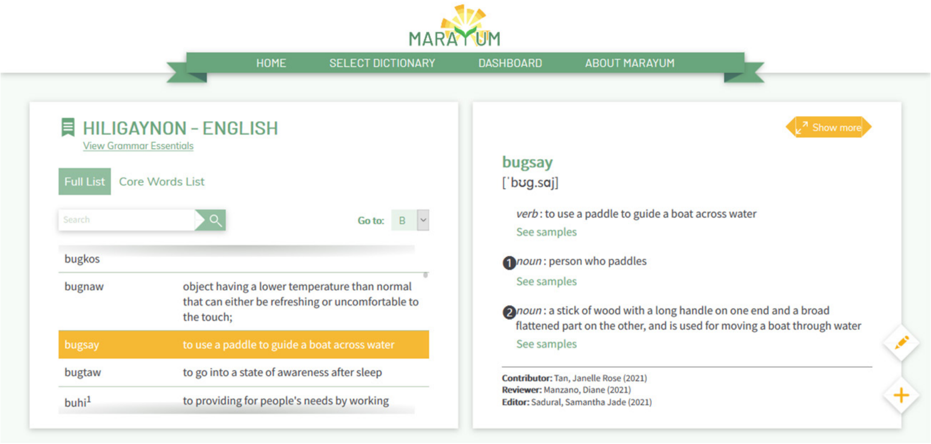
MARAYUM

Empowering the diversity
of Philippine Languages,
one word at a time.

Figure 2: Select Dictionary Popup

The Select Dictionary Popup, as shown in Figure 2, provides a quick way for the user to switch between different dictionaries. This popup is accessible to the user from any screen. The Home screen shows a list of popular dictionaries and a search field for others.

There are two ways to view the contents of the dictionary. These are the Full Word List and the Core Word List.



HILIGAYNON - ENGLISH

View Grammar Essentials

bugkos	
bugnaw	object having a lower temperature than normal that can either be refreshing or uncomfortable to the touch;
bugsay	to use a paddle to guide a boat across water
bugtaw	to go into a state of awareness after sleep
buhi[1]	to providing for people's needs by working

bugsay
[ˈbʊg.sɑj]

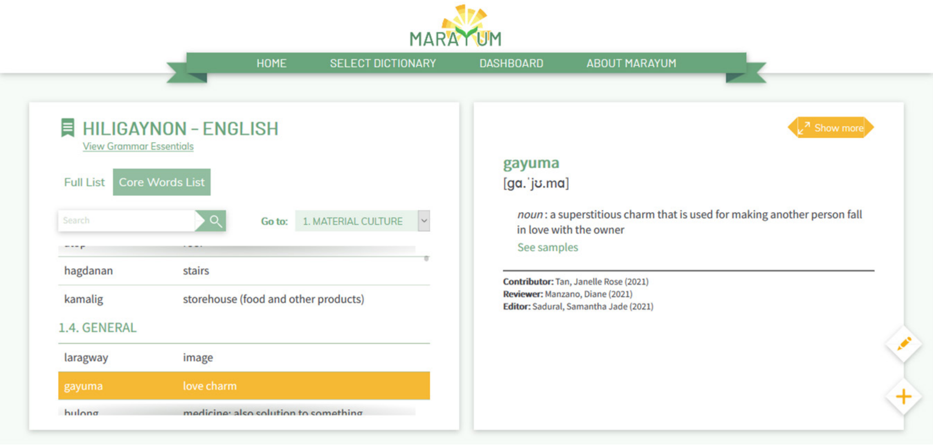
verb : to use a paddle to guide a boat across water
See samples

1 *noun* : person who paddles
See samples

2 *noun* : a stick of wood with a long handle on one end and a broad flattened part on the other, and is used for moving a boat through water
See samples

Contributor: Tan, Janelle Rose (2021)
Reviewer: Manzano, Diane (2021)
Editor: Sadural, Samantha Jade (2021)

Figure 3: Full Word List



HILIGAYNON - ENGLISH

View Grammar Essentials

hagdanan	stairs
kamalig	storehouse (food and other products)

1.4. GENERAL

laragway	image
gayuma	love charm
bulong	medicine; also solution to something

gayuma
[gɑ.ˈjʊ.mɑ]

noun : a superstitious charm that is used for making another person fall in love with the owner
See samples

Contributor: Tan, Janelle Rose (2021)
Reviewer: Manzano, Diane (2021)
Editor: Sadural, Samantha Jade (2021)

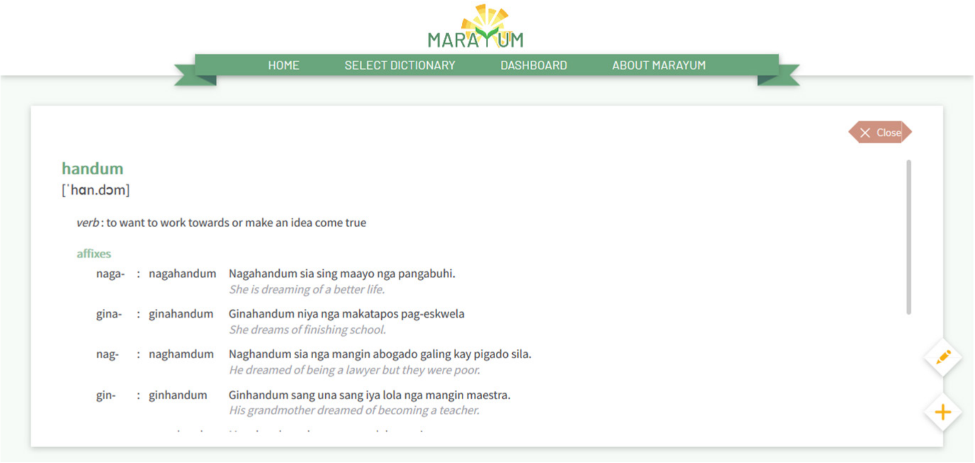
Figure 4: Core Words List

The Full Word List Screen, as shown in Figure 3, shows all the published words in the dictionary at present. This screen will be updated as new words are contributed.

On the other hand, the Core Words List, as shown in Figure 4, shows the core 505 words that form the seed dictionary. These words are predetermined and are constant across languages, which will be discussed in depth in the Marayum Lexicography subsection. These core words are grouped and arranged by semantic domains that describe specific aspects of the language culture of the community that the dictionary belongs to.

The tab buttons for switching between Full List and Core Words List have been simplified to be more intuitive after pre-tests. Both screens also have a search bar to allow for searching for a specific word in the dictionary.

On these screens, the user can click on a word on the list to view its word details on the right side of the screen. Clicking the Show More button will expand the screen into the Word Details Screen.



handum
[ˈhan.dɔm]

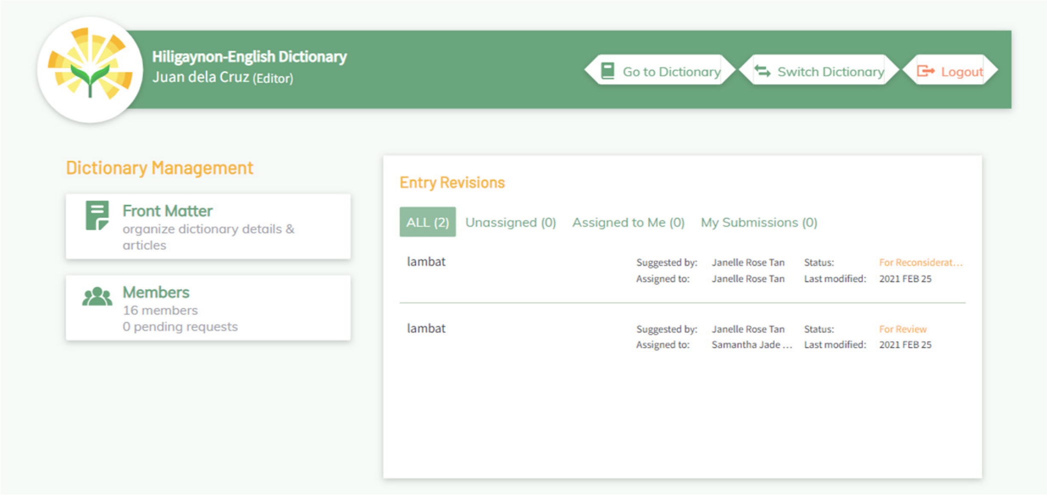
verb : to want to work towards or make an idea come true

affixes

naga-	:	nagahandum	Nagahandum sia sing maayo nga pangabuhi. *She is dreaming of a better life.*
gina-	:	ginahandum	Ginahandum niya nga makatapos pag-eskwela *She dreams of finishing school.*
nag-	:	naghamdum	Naghandum sia nga mangin abogado galing kay pigado sila. *He dreamed of being a lawyer but they were poor.*
gin-	:	ginhandum	Ginhandum sang una sang iya lola nga mangin maestra. *His grandmother dreamed of becoming a teacher.*

Figure 5: Word Details Screen

The Word Details, as shown in Figure 5, panel shows the complete published information about the selected word. This includes the headword, IPA pronunciation, etymology, related words, and definitions if they are available. This screen also shows sample sentences and affixed forms of the selected word entry.



Hiligaynon-English Dictionary
Juan dela Cruz (Editor)

Dictionary Management

Front Matter
organize dictionary details & articles

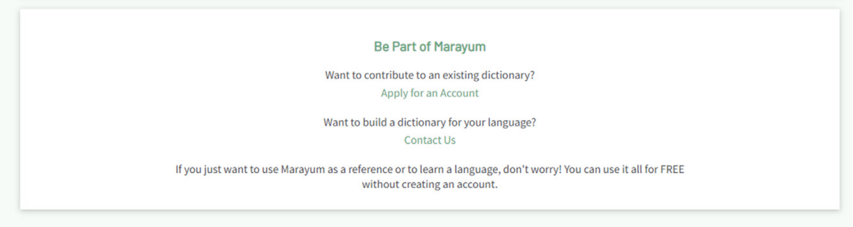
Members
16 members
0 pending requests

Entry Revisions

lambat	Suggested by:	Janelle Rose Tan	Status:	For Reconsiderat...
	Assigned to:	Janelle Rose Tan	Last modified:	2021 FEB 25
lambat	Suggested by:	Janelle Rose Tan	Status:	For Review
	Assigned to:	Samantha Jade ...	Last modified:	2021 FEB 25

Figure 6. Revision Dashboard Screen

The revision system, as shown in Figure 6, is the feature primarily integrated with lexicographic practices. The revision system is only available for authenticated users of the language community. The system will be discussed in detail in the next section. If an unregistered user clicks on the Dashboard, they will be redirected to the Login Screen, as shown in Figure 9, and an invitation to be part of Marayum, as shown in Figure 7.



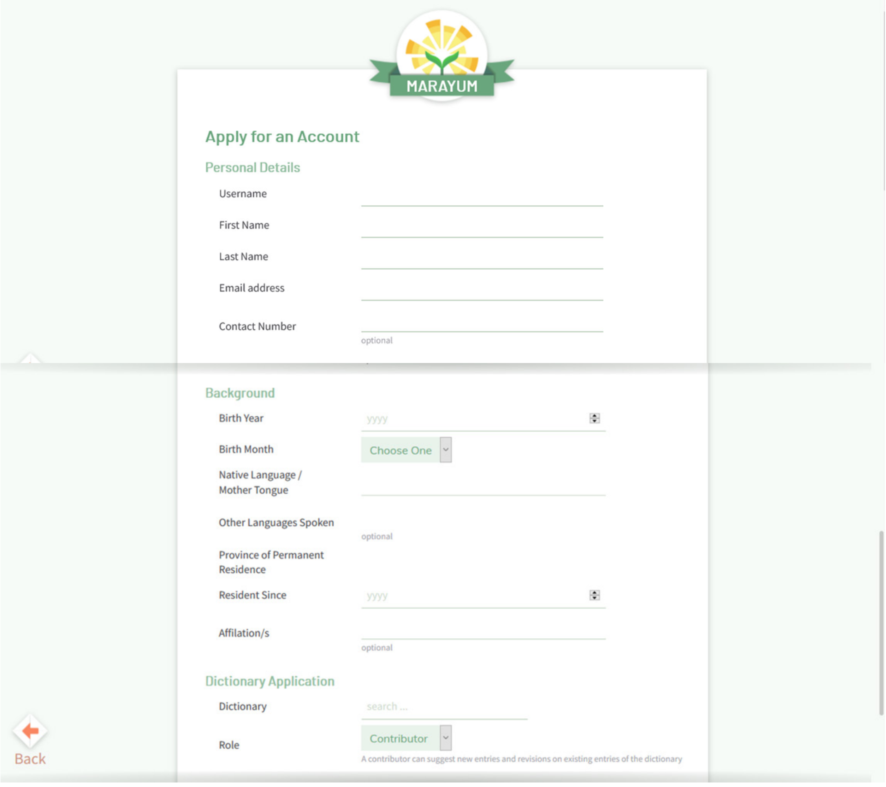
Be Part of Marayum

Want to contribute to an existing dictionary?
Apply for an Account

Want to build a dictionary for your language?
Contact Us

If you just want to use Marayum as a reference or to learn a language, don't worry! You can use it all for FREE without creating an account.

Figure 7: Call for Application Screen



Apply for an Account

Personal Details

Username

First Name

Last Name

Email address

Contact Number
optional

Background

Birth Year
yyyy

Birth Month
Choose One

Native Language / Mother Tongue

Other Languages Spoken
optional

Province of Permanent Residence

Resident Since
yyyy

Affilation/s
optional

Dictionary Application

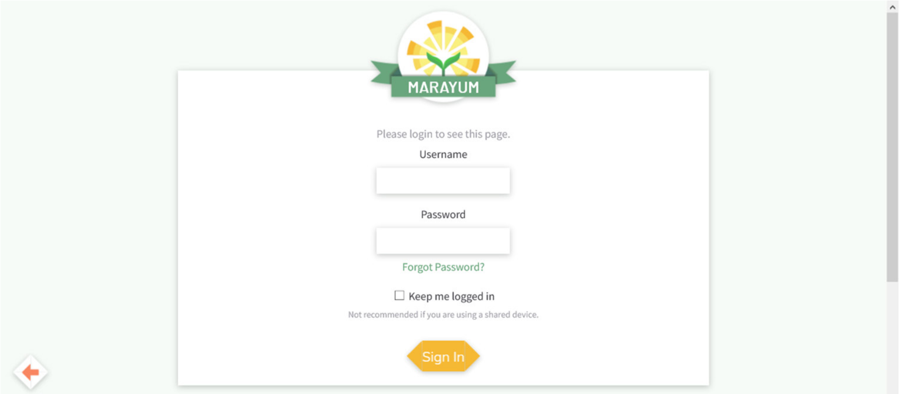
Dictionary
search ...

Role
Contributor
A contributor can suggest new entries and revisions on existing entries of the dictionary

Figure 8A and 8B. Apply for Account Screen

To apply for an account, interested users can provide background information and choose their desired role and dictionary to contribute to. The assigned editors of their chosen dictionary will be responsible for managing their application.The Apply for Account screen is shown in Figure 8A and 8B.

Once their applications are processed, users will receive an email notifying them of the status of their application. Approved users will receive a unique link that will enable them to activate their account and set their password, as seen in Figure 10. Afterward, they can log in to their account and make contributions. Figure 9 shows the Login screen.



Please login to see this page.

Username

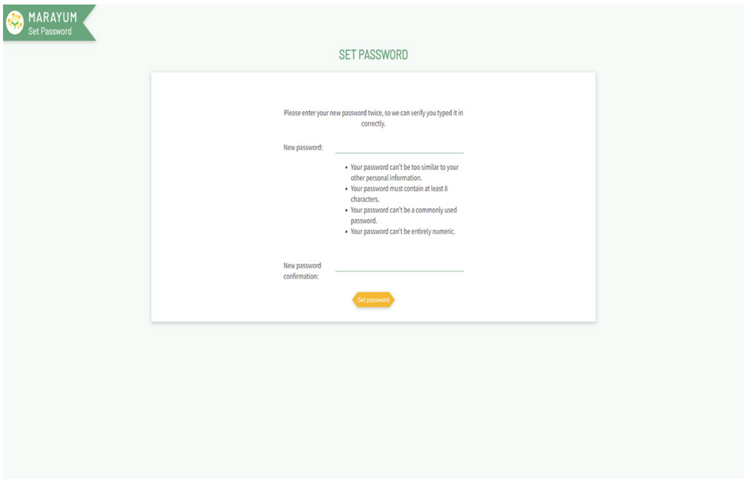
Password

□ Keep me logged in

Not recommended if you are using a shared device.

Figure 9. Login Screen

As seen in Figure 9, a “Keep me logged in” checkbox has been added to consider users on shared devices. If the checkbox is not checked on login, the user is automatically logged out once the browser is closed or the device is turned off.



Please enter your new password twice, so we can verify you typed it in correctly.

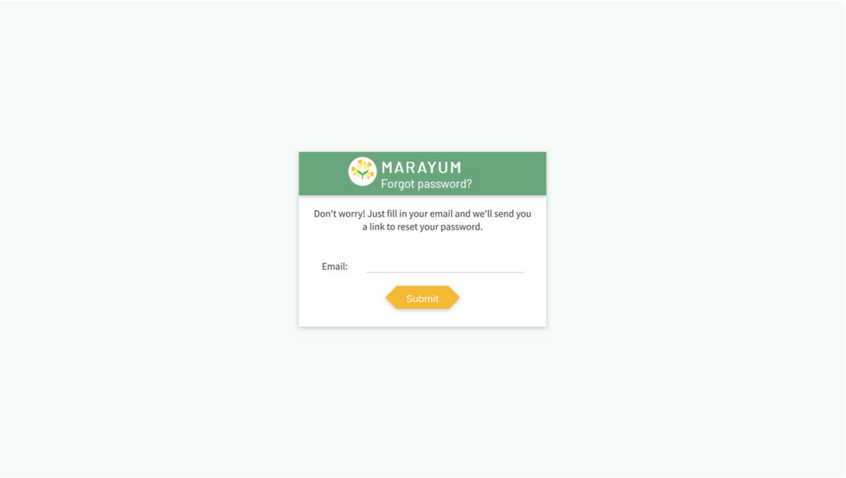
New password:

- Your password can’t be too similar to your other personal information.
- Your password must contain at least 8 characters.
- Your password can’t be a commonly used password.
- Your password can’t be entirely numeric.

New password confirmation:

Figure 10. Set Password Screen

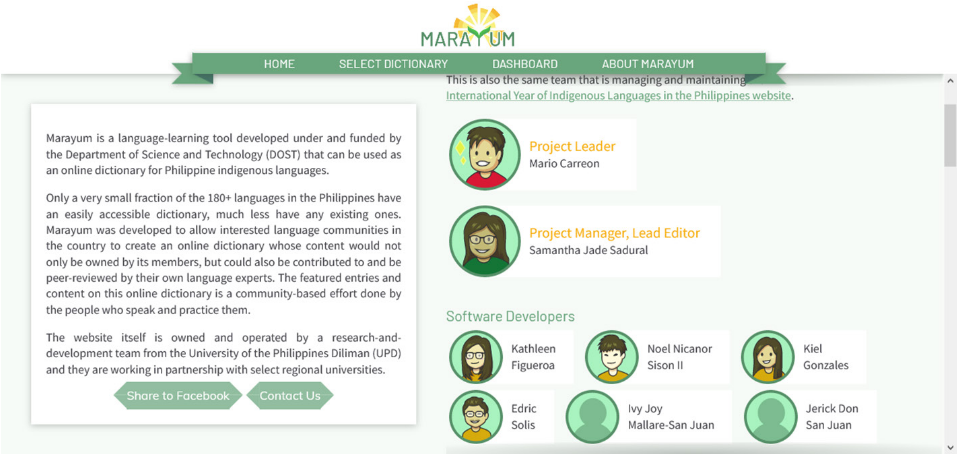
When a user forgets their password, they can initiate a password reset by submitting their email, as shown in Figure 11. A password reset link will be sent to their email address.



Don’t worry! Just fill in your email and we’ll send you a link to reset your password.

Email:

Figure 11. Password Reset Screen



Marayum is a language-learning tool developed under and funded by the Department of Science and Technology (DOST) that can be used as an online dictionary for Philippine indigenous languages.

Only a very small fraction of the 180+ languages in the Philippines have an easily accessible dictionary, much less have any existing ones. Marayum was developed to allow interested language communities in the country to create an online dictionary whose content would not only be owned by its members, but could also be contributed to and be peer-reviewed by their own language experts. The featured entries and content on this online dictionary is a community-based effort done by the people who speak and practice them.

The website itself is owned and operated by a research-and-development team from the University of the Philippines Diliman (UPD) and they are working in partnership with select regional universities.

This is also the same team that is managing and maintaining International Year of Indigenous Languages in the Philippines website.

Project Leader
Mario Carreon

Project Manager, Lead Editor
Samantha Jade Sadural

Software Developers

Kathleen Figueroa
Noel Nicanor Sison II
Kiel Gonzales
Edric Solis
Ivy Joy Mallare-San Juan
Jerick Don San Juan

Figure 12. About Screen

Figure 11 shows the About page, a screen that showcases information about the project, the team, and the institutions that took part in bringing the project to communities.

4 Analysis and Discussion

4.1 Overview: The Lexicography of Marayum

Table 1 provides a summary of a few website terminologies essential to the discussion of this section.

Table 1: Marayum Website Terminologies

Term	**Description**
Revision	Data of a word. Contains information such as headword, language, pronunciation, variants, derivatives, descriptions, and example sentences.
Contributor	Logged in users who are permitted to submit new Revisions. They can also edit their submitted revisions as long as it is not yet under review by a Reviewer or Editor.
Reviewer	Can list, view, and edit Revisions submitted by Contributors. After reviewing a Revision, they will mark it as "For Publishing," "For Rejection," and "For Editor."
Editor	Can list, view, and edit Revisions submitted by Contributors and reviewed by Reviewers. After reviewing a revision, they can either publish the word in the dictionary or reject it.
Base Language	The language of a word. Contributor, Reviewer, and Editor roles are assigned to a specific language and can only submit or review Revisions for that language. A user may be granted more than one role for one or more languages.

Weigand's (1998, as cited in Schierholz, 2015) practical lexicographic approach is valid for both print and online dictionaries. The phases, integrated with documentary lexicography, are followed in the production of Marayum dictionaries—may it be a completely new project, a dictionary derived from other existing dictionaries, or a revision of an existing dictionary. Below is the list of the practical lexicographic approach

1. Preparation and planning phase
2. Material collection phase
3. Material processing phase
4. Material evaluation phase
5. Publication preparation phase
6. Publication phase

4.2 Preparation and Planning Phase: Marayum's Software Architecture

The preparation and planning phase of the practical lexicographic process is geared towards the actual cost of the project, the workflow, the agreed length of making the project, the size of the dictionary, as well as the work schedule of the lexicographic work and the distribution of the tasks (Schierholz, 2015). As the majority of these processes are at the administrative level of managing the dictionary, this process was actualized in the software architecture of Marayum.

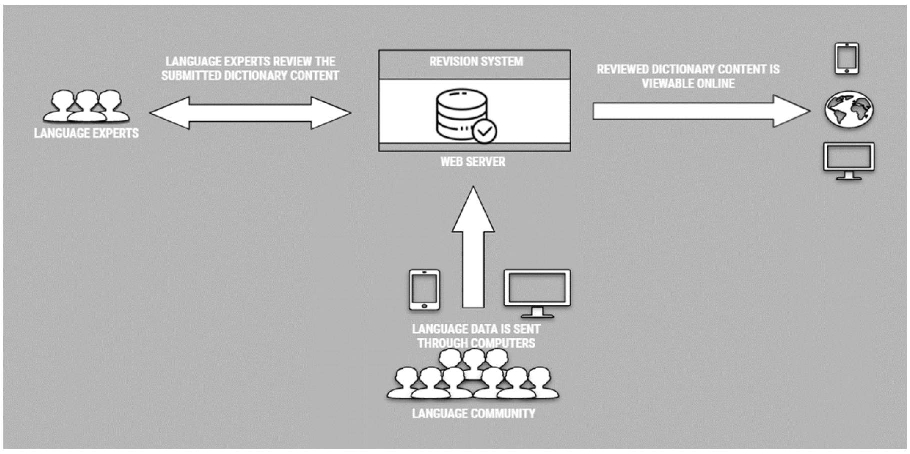

LANGUAGE EXPERTS REVIEW THE SUBMITTED DICTIONARY CONTENT
REVISION SYSTEM
REVIEWED DICTIONARY CONTENT IS VIEWABLE ONLINE
LANGUAGE EXPERTS
WEB SERVER
LANGUAGE DATA IS SENT THROUGH COMPUTERS
LANGUAGE COMMUNITY

Figure 13. Marayum's System Architecture

Project Marayum's System Architecture, as shown in Figure 13, was designed to implement the practical lexicographic approach integrated with documentary lexicography practices. The system architecture was designed to allow a language community to have an online dictionary without having the technical background or expertise necessary to have an online presence.

4.2.1 Administrative Tasks

The community will be responsible for creating the language dictionary and contributing more content to this system over time. The users who applied as Contributors in Marayum can suggest more words to the dictionary, propose changes to existing words, provide a sample or illustrative sentences, record speech samples, and in time, add richer content to the corpus. The only requirement is that the Contributor is a native speaker of the language.

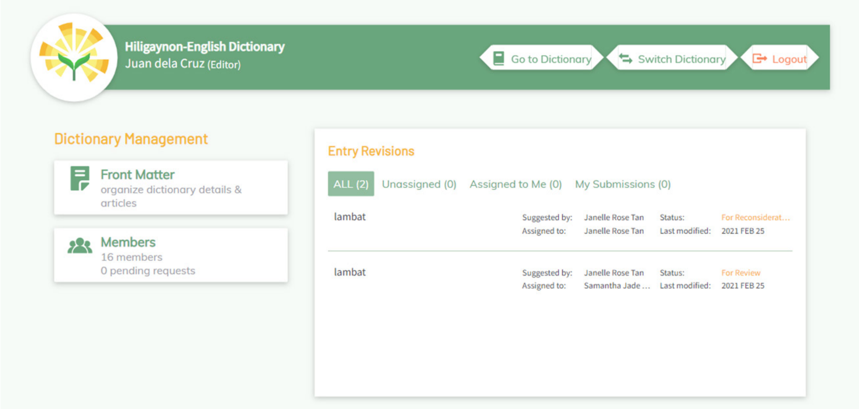
These contributions are reviewed by a panel of language experts to ensure the quality of the dictionary being created. These language experts, labeled as Reviewers in Marayum, are required to have language education or linguistic training. The Reviewers are responsible for the appropriateness of the content and the assignment of certain lexicographic terms, e.g., part of speech, IPA representation of the word, and additional grammatical information.

New content approved by the panel is then published online. The Editor of the language dictionary is given to the linguist or language expert responsible for the documentation and preservation efforts of the community. The setup and maintenance of the revision system are handled by the Project Marayum team, allowing the language communities to focus on their dictionary's content.

The administrative task and maintenance of the language community's dictionary will be assigned to the state university or non-government organization spearheading the language preservation efforts of the community. The editor, university, and other partners should work hand-in-hand in assuring the sustainability of their language dictionary.

4.2.2 Revision System

The Marayum server contains the website hosting the language dictionaries and its content management system, called the Revision system. Upon logging in to Marayum, the Revision Dashboard screen will be shown, as seen in Figure 14.

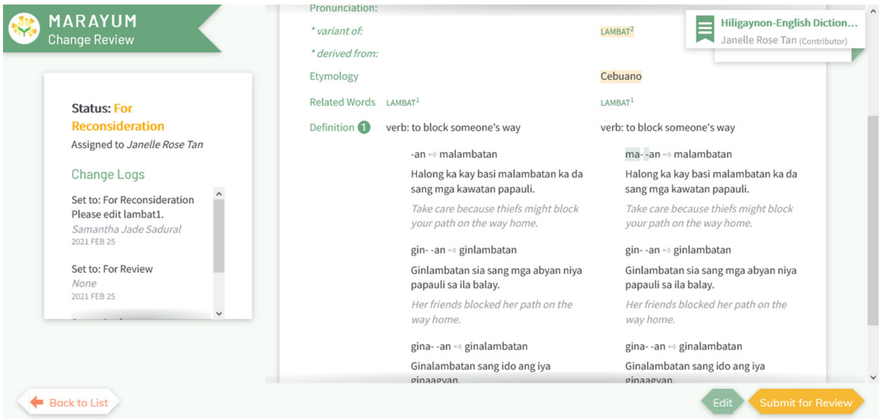


lambat
Suggested by: Janelle Rose Tan Status: For Reconsiderat...
Assigned to: Janelle Rose Tan Last modified: 2021 FEB 25

lambat
Suggested by: Janelle Rose Tan Status: For Review
Assigned to: Samantha Jade ... Last modified: 2021 FEB 25

Figure 14. Revision Dashboard Screen

Only registered members of the language community are allowed access to this screen. In addition to access to screens available to the general public, the list of revisions is also displayed according to the role of the user. Contributors can only see revisions that they submitted. Reviewers and Editors can see unassigned and their assigned revisions. Figure 15 shows an example of the Revision Review Screen.



Status: For Reconsideration
Assigned to *Janelle Rose Tan*

Change Logs

Set to: For Reconsideration
Please edit lambat1.
Samantha Jade Sadural
2021 FEB 25

Set to: For Review
None
2021 FEB 25

Pronunciation:
** variant of:* LAMBAT2
** derived from:*
Etymology: Cebuano
Related Words: LAMBAT1 | LAMBAT1
Definition 1: verb: to block someone's way | verb: to block someone's way

-an ⇨ malambatan	ma- -an ⇨ malambatan
Halong ka kay basi malambatan ka da sang mga kawatan papauli.	Halong ka kay basi malambatan ka da sang mga kawatan papauli.
Take care because thiefs might block your path on the way home.	*Take care because thiefs might block your path on the way home.*
gin- -an ⇨ ginlambatan	gin- -an ⇨ ginlambatan
Ginlambatan sia sang mga abyan niya papauli sa ila balay.	Ginlambatan sia sang mga abyan niya papauli sa ila balay.
Her friends blocked her path on the way home.	*Her friends blocked her path on the way home.*
gina- -an ⇨ ginalambatan	gina- -an ⇨ ginalambatan
Ginalambatan sang ido ang iya ginaagyan.	Ginalambatan sang ido ang iya ginaagyan.

Figure 15. Revision Review Screen

Users, as seen in the Revision Review Screen (Figure 15), can perform selected actions on the revision based on their role and the current revision stage. A reviewer may return a revision to a contributor, pass it to an editor with recommendations, or reject it. An editor may return the revision to a reviewer or publish the revision to update the original dictionary entry. Figure 7 shows the steps a revision can go through.

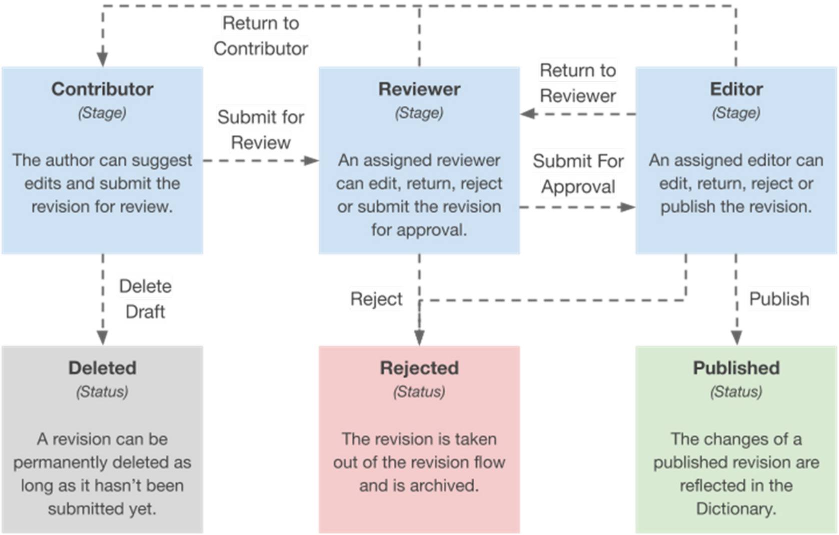

Return to Contributor
Return to Reviewer
Contributor
(Stage)
The author can suggest edits and submit the revision for review.
Submit for Review
Reviewer
(Stage)
An assigned reviewer can edit, return, reject or submit the revision for approval.
Submit For Approval
Editor
(Stage)
An assigned editor can edit, return, reject or publish the revision.
Delete Draft
Reject
Publish
Deleted
(Status)
A revision can be permanently deleted as long as it hasn't been submitted yet.
Rejected
(Status)
The revision is taken out of the revision flow and is archived.
Published
(Status)
The changes of a published revision are reflected in the Dictionary.

Figure 16. Stages of a Revision

The buttons displayed on the bottom right area of the Revision Review Screen (Figure 15) change depending on the revision's stage and the currently logged-in user's role. The Edit button is only available to the user assigned at the revision's current stage. The buttons are no longer available when the revision is finalized, such as Deleted, Rejected, and Published.

To help with the coordination between the different users, a Change Review log, seen on the left side on Figure 15), shows the status changes and comments made as the word entry undergoes revision. For example, when a user attempts to change the revision's stage and another user is assigned at the target stage, the initiating user will be required to enter a brief comment for context.

Contributors can suggest new lexemes for the dictionary through the Create or Edit Lexeme Screens, as shown in Figures 17 and 18. The user only needs to enter the word, pronunciation, part of speech, and a definition for at least one sense. The remaining fields, such as etymology, related words, and example sentences, are optional. As mentioned earlier, the revisions created by a user will have to go through a reviewer and editor for approval before it can be published to the public.

4.2.3 Lexicographic Information

Marayum follows the structure of a bilingual dictionary, a Philippine-type language as its L1 and English as its L2. It is also intended to be a general-language dictionary; in the long run, however, the seed dictionary version opted for an onomasiological structure following the 505- word list categories.

Compared to a specialized dictionary, a general-language dictionary can cover the breadth of the language and its culture. As one of the project's aims is language documentation, this format is more apt. This method also attempts to collate the language's sample sentences, which can be used as a preliminary dataset to create that language's corpus.

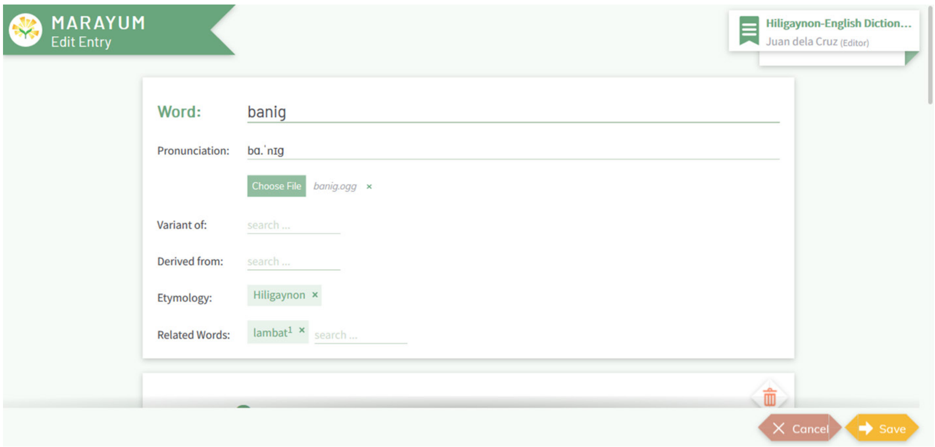
As Marayum is based on documentary lexicography, the project also aims to aid in language documentation, especially endangered languages. The Philippines is a language hotspot, which means that it is an area with many languages near extinction. Asi (bno), one of the languages with an initial dictionary in Marayum, while categorized as a developing language in the Ethnologue, is unstandardized and not used in formal situations. In the long run, the tool can ease the process of documenting even moribund to nearly extinct languages.

It can also serve as an archive for these present and future efforts. Digital language archiving is also an aspect of linguistic research that has been overlooked. To date, there are no local digital data archives available for free, much less a corpora specific for linguistic, sociological, and anthropological research.

In line with this, the seed dictionary consisting of 505 words following an onomasiological format is consistent with the project's aims towards language and lexicographic documentation. An onomasiological dictionary groups words instead of the usual a-z index that most dictionaries use. Onomasiology, a branch of linguistics, is primarily concerned with the question "How do you express X?" which can help draw out the meaning of a particular word. Usually, this process can push the user to categorize the word before saying the word's meaning equivalent, making the elicitation process more conscious. The seed dictionary also utilizes the elicitation of sample sentences, which can further aid a linguist in defining a certain word based on its use grammatically and in accordance with other elements and arguments used in the sentence.

Initial lemma selection for the seed dictionary is already pre-chosen, removing the burden of having the community decide on the initial words for their dictionary. It was, however, highly suggested that the contributors associate related lexemes and add them to the dictionary. This strategy was proven effective; for example, Marayum Kinaray-a writers finished the seed dictionary with 2,426 words from the original 505 words. The dictionary also contains Kinaray- a sentences with corresponding English translations. For other languages, the seed dictionary for Hiligaynon now contains 1,789 words, which also contains Hiligaynon sentences, as well as its subsequent English translations. The Cebuano (Southern Leyte variant) writers have also finished the seed dictionary, with the 505 words growing to 1,120 words, while Asi writers expanded their initial words to 1,130 words.

Creating new words or lexemes, as well as editing existing words, in Marayum was streamlined and designed to be user-friendly, as seen in Figures 17 and 18.



Word: banig

Pronunciation: bɑ.ˈnɪg

banig.ogg

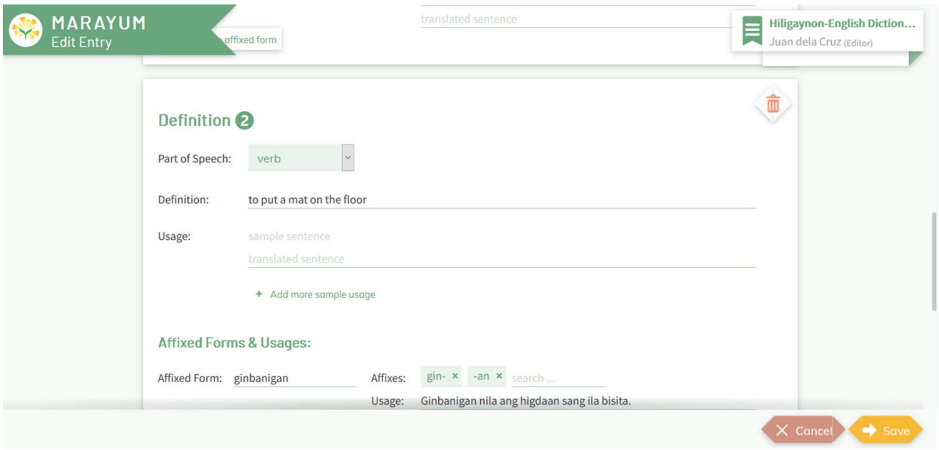
Variant of:

Derived from:

Etymology: Hiligaynon

Related Words: lambat[1]

Figure 17. Create or Edit Lexeme Screens Part 1



Definition 2

Part of Speech: verb

Definition: to put a mat on the floor

Usage:

Affixed Forms & Usages:

Affixed Form: ginbanigan

Affixes: gin- -an

Usage: Ginbanigan nila ang higdaan sang ila bisita.

Figure 18. Create or Edit Lexeme Screens Part 2

The following lexicographic information can be encoded by the contributor when creating a lexeme or editing an existing one.

- Headword
- Pronunciation
- Etymology
- Derived From
- Variant Of

Additionally, this form contains one or more sections for the Definition fields. Each section contains the following field groups: Part of Speech, Definition, and Usage. The Definition field is a Text Field for the word's definition in the second language.

Usage fields can have one or more rows where each row has an Affix, Affixed Form, Sample Sentence, and Translated Sentence. Affix and Affixed Form are optional fields. Affix is a dropdown list of affixes in the base language and an option to add a new affix. If the latter option is selected, a new Affix can be entered in a separate Text Field. Affixed Form is a simple Text Field where the user can choose the form of the word based on the selected affix. The last 2 Usage fields are Text Fields for Example or Illustrative Sentences in the base language and second language.

The revisions created by a user will have to go through a reviewer and editor for approval before it can be published to the public.

4.3 Material Collection Phase: Documentary Elicitation Methods

Material collection is based on the experience and knowledge of the project staff of a given language dictionary. The documentation practices of the University of the Philippines Diliman Department of Linguistics and the community documentary efforts of the Asi Studies Center for Culture and Arts (ASCCA) were then merged into the roles in Marayum, as shown in Figure 16. However, Marayum is a tool, and the elicitation and collection practices of the groups are as follows.

4.3.1 Language Community Efforts and Roles

ASCCA, a non-profit non-government organization spearheading the documentation and preservation of Asi/Bantoanon culture, arts, and language, created a team focusing on the creation of their own language dictionary. The contributors, reviewers, and editors are part of the community, with the reviewer and editor roles filled by teachers and linguists alike.

The elicitation process, in this case, follows the straightforward elicitation design of Marayum, as seen in Figure 16.

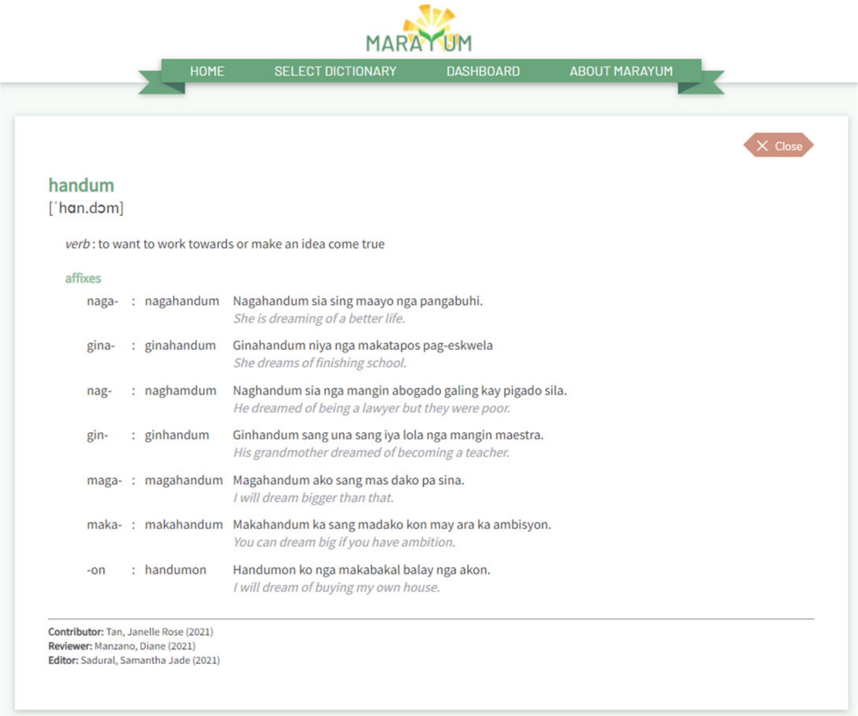
4.3.2 University/Department-led Fieldwork Elicitations

Fieldwork elicitations led by university lexicography and linguistics professors and implemented by language or linguistics students are also considered. The state of network infrastructure in the Philippines was also reviewed, as there is a high chance that communities with endangered languages are situated in places with low or no internet connectivity.

In this case, the contributors (with their explicit consent) will still be the native language speakers of the community, with the Reviewers being the students or linguists collating the data and doing the fieldwork. The Editor can be the assigned linguist of the language community or the linguistic or language professor of the reviewers.

4.3.3 Credits and Sources

Marayum's word page, as seen in Figure 19, has a Credits Footer to acknowledge the language community and its registered contributors, reviewers, and editors. Doing so seals the ownership of the community over their dictionary.



handum
[ˈhɑn.dɔm]

verb : to want to work towards or make an idea come true

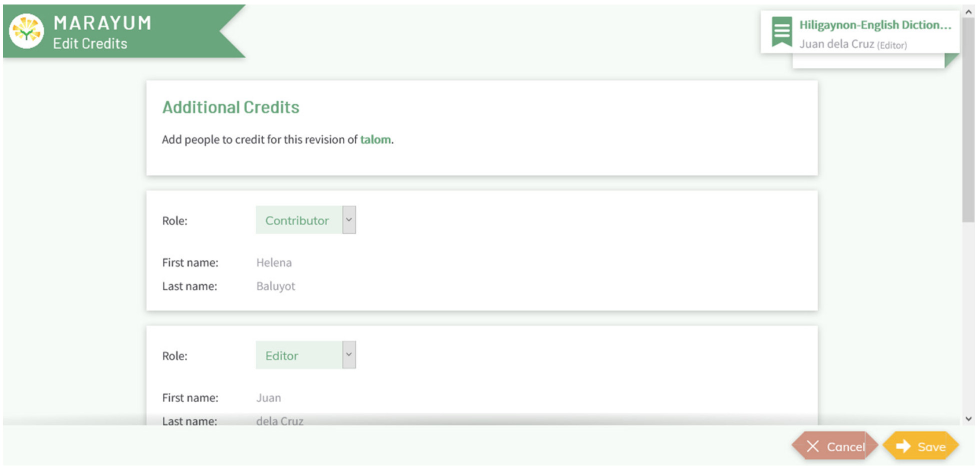
affixes

naga-	:	nagahandum	Nagahandum sia sing maayo nga pangabuhi. *She is dreaming of a better life.*
gina-	:	ginahandum	Ginahandum niya nga makatapos pag-eskwela *She dreams of finishing school.*
nag-	:	naghamdum	Naghandum sia nga mangin abogado galing kay pigado sila. *He dreamed of being a lawyer but they were poor.*
gin-	:	ginhandum	Ginhandum sang una sang iya lola nga mangin maestra. *His grandmother dreamed of becoming a teacher.*
maga-	:	magahandum	Magahandum ako sang mas dako pa sina. *I will dream bigger than that.*
maka-	:	makahandum	Makahandum ka sang madako kon may ara ka ambisyon. *You can dream big if you have ambition.*
-on	:	handumon	Handumon ko nga makabakal balay nga akon. *I will dream of buying my own house.*

Contributor: Tan, Janelle Rose (2021)
Reviewer: Manzano, Diane (2021)
Editor: Sadural, Samantha Jade (2021)

Figure 19: Word Page and Credits Footer

To consider workflows where Reviewers and Editors collect field data offline, additional credits for a revision can be added at a separate screen, as shown in Figure 20.



Additional Credits

Add people to credit for this revision of talom.

Role: Contributor
First name: Helena
Last name: Baluyot

Role: Editor
First name: Juan
Last name: dela Cruz

Figure 20: Edit Credits Screen

As of this writing, Marayum can only handle primary sources collated through dictionary writing. Future versions of Marayum will explore secondary and tertiary sources for collection.

4.4 Material processing phase: NLP techniques on low data-resource language

In this day and age, methods on this phase expect big electronic data and natural language processing techniques for lemma candidates, word frequencies, and collocation candidates, among others. However, due to the lack of corpora in Philippine languages, the 505 wordlist served as the lemma selection primer for the development of seed dictionaries in Marayum.

An initial study on building a corpus of Asi/Bantoanon language, a low resource language, was undertaken alongside the development of Marayum, wherein the Marayum team sought to identify additional sources of Asi literature. Unfortunately, both Romblon State University and ASCCA do not have an archive of Asi works, nor do they have an Asi corpus on hand.

Most of the books being transcribed by the team are the personal works and collections of the head of ASCCA, Ismael Fabicon. The team also handled the encoding of these into a digital format, assisted by optical character recognition software.

All in all, approximately 107,599 words were gathered. Of these, there are 31,163 unique words. This small database was used to perform an initial analysis of Asi word frequency count, frequency drop-off, and bigram language model. Table 2 below shows the top 20 most used words in Asi and how often they appear in the corpus.

Table 2: Asi Word Frequency Count

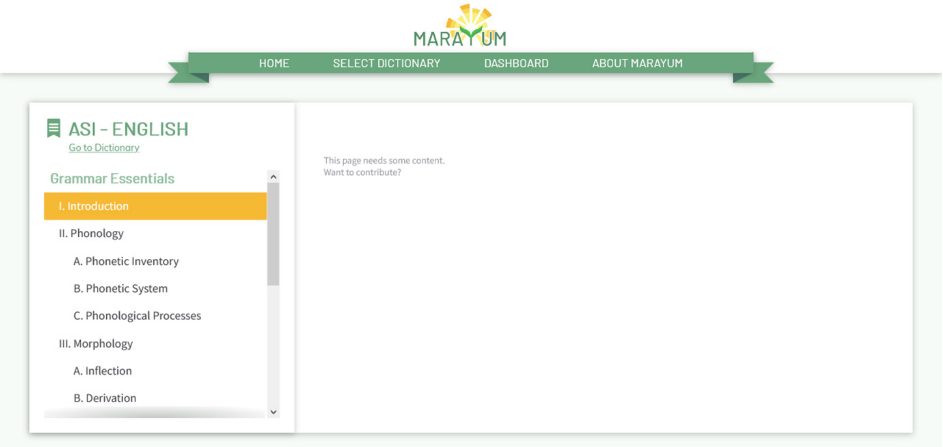
0.493%	ATO
0.449%	WAYA
0.442%	NA
0.431%	SI
0.428%	YANG
0.417%	IDA

The most common word in the corpus is SA which takes up 5.8% of the entire corpus. This is followed by IT, KA, NAK, which take up 3.90%, 3.23%, and 3.19%, respectively.

A planned feature, the Corpus Management System, was to mitigate the low word count of Asi by allowing users to upload literature samples even after project implementation has ended. The scripts to generate NLP metrics from new content do exist but are not integrated with the Marayum website. New Asi literature, however, can still be added into the corpus, with the metrics generated from these scripts. Additional languages can also benefit from this feature.

4.4 Material evaluation phase: Dictionary's Front Matter

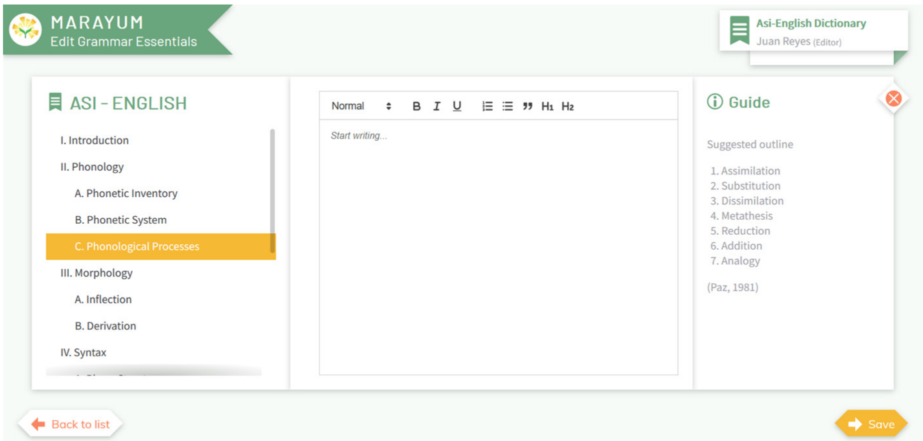
This phase is where the dictionary articles are drawn up based on the collected and prepared material (Schierholz, 2015). The philological method set by the 505 words, for now, is being followed by Marayum, being based on primary sources written by native speakers of the language. Additional articles are collated in the Front Matter section of the dictionary.



This page needs some content.
Want to contribute?

Figure 21: Front Matter List Screen

As shown in Figure 21, the Front Matter List Screen displays a list of articles that users can view for more in-depth information about a language. Clicking on an article title on the left- side outline will display the article contents on the right-side panel.



Suggested outline

1. Assimilation
2. Substitution
3. Dissimilation
4. Metathesis
5. Reduction
6. Addition
7. Analogy

(Paz, 1981)

Figure 22: Front Matter Editor Screen

Users can change Front Matter information with the Editor role. Formatting text is supported using the Quill Rich Text Editor. Some sections have writing suggestions displayed on the Guide panel beside the text editor.

4.5 Publication preparation and publication phase: Reviewer and Editor Roles

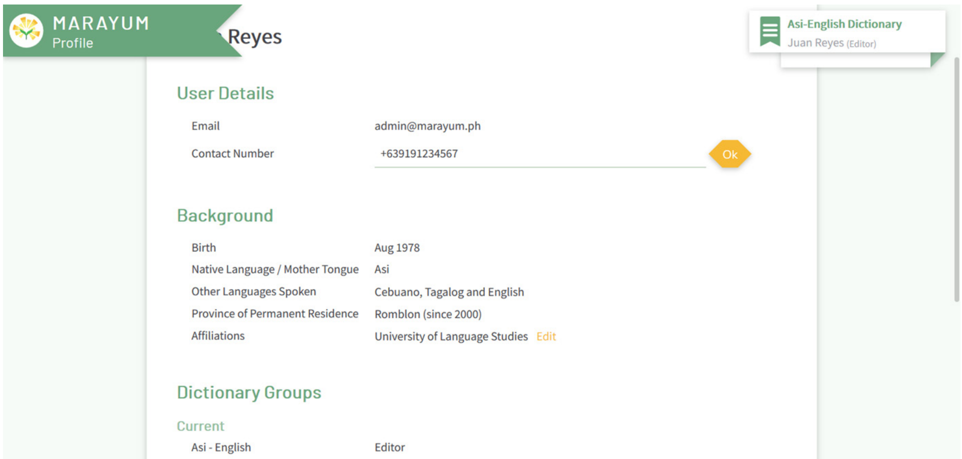
Traditionally, the lexicographer has editorial control of all entries and articles in a dictionary (Schierholz, 2015, Svensen, 2009). In Marayum, to carry these processes out reliably, lexemes undergo two levels of review, that of the Reviewer and the Editor before publication. The usage of Marayum itself for creating dictionaries streamlines both the publication preparation and publication phase of the project.

A dictionary in Marayum is published on an online dictionary medium. It will follow the standard link https://marayum.ph/dictionary/(language)-english/. All dictionaries will be a dynamic dictionary, with the community involved in updating their dictionary by suggesting new words, editing old ones, and adding or editing sample sentences online. Updated versions will be published under the discretion of the language editor of the dictionary.

4.6 Data maintenance and Post-production

The UP Department of Computer Science will handle the technical maintenance of the output of Project Marayum. The team will also work hand in hand with UP OVCRD to discuss and implement the project's sustainability in collaboration with the UP Diliman Department of Linguistics.

Maintenance of the dictionary is important for a dynamic dictionary. User management is particularly relevant for Marayum, given the regular influx of users willing to contribute to their dictionary.



User Details

Email	admin@marayum.ph
Contact Number	+639191234567

Background

Birth	Aug 1978
Native Language / Mother Tongue	Asi
Other Languages Spoken	Cebuano, Tagalog and English
Province of Permanent Residence	Romblon (since 2000)
Affiliations	University of Language Studies

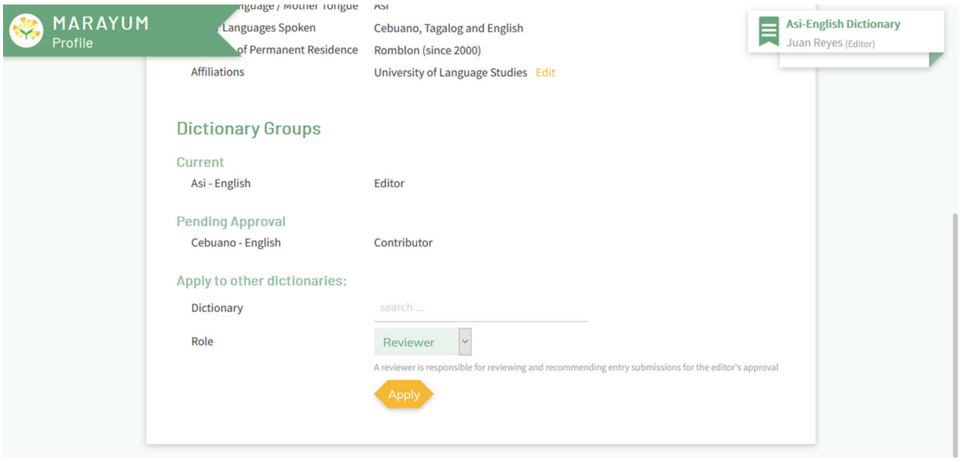
Dictionary Groups

Current

Asi - English	Editor

Figure 23: Profile Screen: Background Information

The currently logged-in user can view the information they submitted during the account application in the Profile Screen, as seen in Figure 23. Hovering over the information text displays an "Edit" link that can be clicked to display an inline editing form. Once the user clicks the "OK" button, the changes are saved.



nguage / Mother Tongue	Asi
Languages Spoken	Cebuano, Tagalog and English
of Permanent Residence	Romblon (since 2000)
Affiliations	University of Language Studies

Dictionary Groups

Current

Asi - English	Editor

Pending Approval

Cebuano - English	Contributor

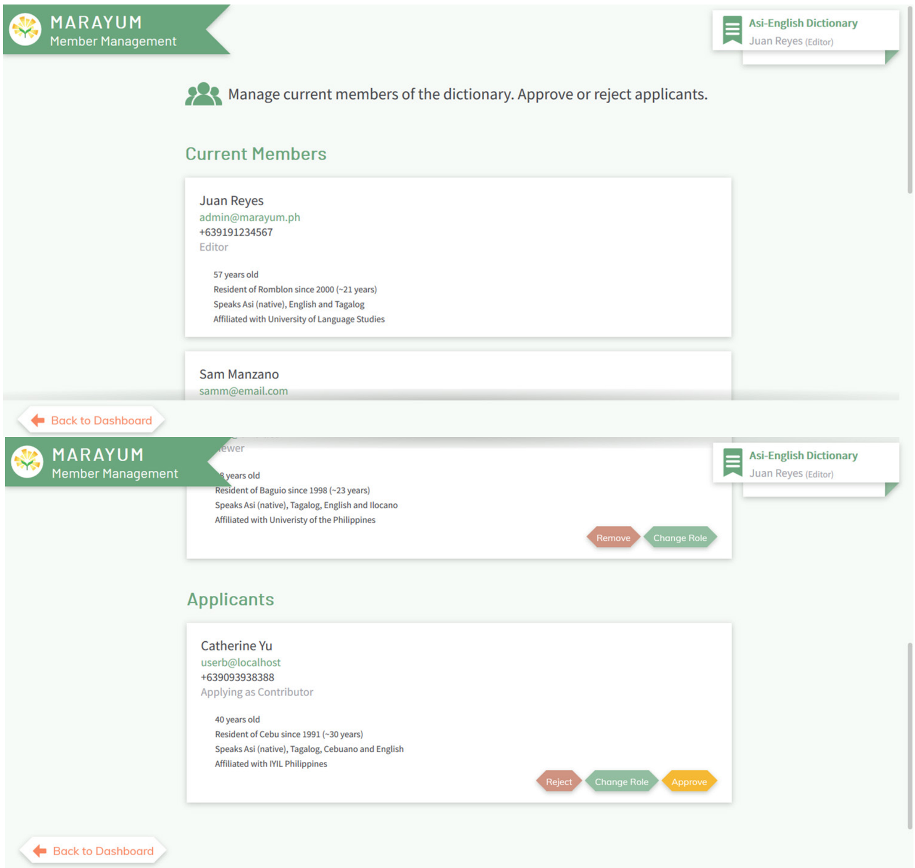
Apply to other dictionaries:

Dictionary	
Role	Reviewer

A reviewer is responsible for reviewing and recommending entry submissions for the editor's approval

Figure 24: Profile Screen: Dictionary Roles

As seen in Figure 24, the bottom half of the Profile Screen shows their dictionary roles, dictionaries they have applied for, and a form that allows them to apply for more dictionary roles. When a user selects a role using the dropdown, a short description of the role appears below it.



Manage current members of the dictionary. Approve or reject applicants.

Current Members

Juan Reyes
admin@marayum.ph
+639191234567
Editor

57 years old
Resident of Romblon since 2000 (~21 years)
Speaks Asi (native), English and Tagalog
Affiliated with University of Language Studies

Sam Manzano
samm@email.com

ewer

8 years old
Resident of Baguio since 1998 (~23 years)
Speaks Asi (native), Tagalog, English and Ilocano
Affiliated with Univeristy of the Philippines

Applicants

Catherine Yu
userb@localhost
+639093938388
Applying as Contributor

40 years old
Resident of Cebu since 1991 (~30 years)
Speaks Asi (native), Tagalog, Cebuano and English
Affiliated with IYIL Philippines

Figures 25A & 25B: Manage Users Screen

Users can access the Manage Users Screen to view the list of co-members in the dictionary. Additionally, editors can view the background information and edit roles of existing members. They can also accept and reject applications from new users who want to contribute to the dictionary. Users are sent an email when their role or application is updated.

5 Conclusion

Although this paper has focused on creating language dictionaries without utilizing language corpora, Project Marayum has made some efforts to create a corpus of the Asi (bno) language. Through the course of project implementation, approximately 107,599 words were gathered. Of these, there are 31,163 unique words.

To get around the low resource limitation, Project Marayum has an initial Corpus Management System. Its goal is to allow users to upload literature samples even after project implementation has ended. These literature samples will be added to its language corpus. Top 100 commonly- used words not in the existing Marayum dictionary will be prioritized in the lemma selection. It is envisioned that as the corpus grows, the dictionary will grow alongside it.

However, due to project time restrictions and the rise of the Novel Coronavirus pandemic, the Corpus Management system is shelved and will be developed at a later date.

Alongside additional development plans for Marayum, suggestions on improving the lexicography of Marayum is highly encouraged. While Marayum is not the first online dictionary made for Philippine languages, it is the first of its kind to consider direct community involvement, giving the users their right to write, edit, review, and own their dictionary.

5 References

Crystal, D. (2002). *Language Death.* Cambridge: Cambridge University Press.

Dita, S., Roxas, R. E., & Inventado, P. (2009, December). Building online corpora of philippine languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2* (pp. 646-653).

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). (2021) *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International.

Department of Education. (2012). *Guidelines of the Implementation of the Mother Tongue- based Multilingual Education (MTB-MLE).* (DO 16, S. 2012). https://www.deped.gov.ph/2012/02/17/do-16-s-2012-guidelines-on-the- implementation-of-the-mother-tongue-based-multilingual-education-mtb-mle/

Department of Education. (2013a). *K to 12 Mother Tongue Curriculum Guide*. https://www.deped.gov.ph/wp-content/uploads/2019/01/Mother-Tongue-CG.pdf

Department of Education. (2013b). *Additional guidelines to DepEd Order No. 16, S. 2012 (Guidelines of the Implementation of the Mother Tongue-based Multilingual Education (MTB-MLE)).* (DO 28, S. 2013). https://www.deped.gov.ph/2013/07/05/do-28-s-2013 -additional-guidelines-to-deped-order-no-16-s-2012-guidelines-on-the- implementation-of-the-mother-tongue-based-multilingual-education-mtb-mle/ RetrievedonApr2021.

Department of Education. (2020). *DepEd open to more dialogue on improvement of MTB-MLE implementation.* https://www.deped.gov.ph/2020/02/28/deped-open-to-more- dialogue-on-improvement-of-mtb-mle-implementation/

Hacker, K. (2013). *Community-based participatory research*. Sage publications.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Lee, A. P. (2010). The Filipino monolingual dictionaries and the development of Filipino lexicography. *Philippine Social Sciences Review*, *62*(2).

Lew, R. (2013). *User-generated content (UGC) in online English dictionaries.* OPAL: Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache.

Lewis, M. P., & Simons, G. F. (2010). Assessing endangerment: expanding Fishman's GIDS. *Revue roumaine de linguistique*, *55*(2), 103-120.

Schierholz, S. J. (2015). Methods in lexicography and dictionary research. *Lexikos*, *25*, 323- 352.

Reysio-Cruz, M., (2019, August 19). Saving PH diverse languages from extinction. *Philippine Daily Inquirer.* https://newsinfo.inquirer.net/1155014/saving-ph-diverse- languages- from-extinction.

Rundell, M., Hanks, P., & Schryver, G. M. D. (2015). Crowdsourcing, wikis, and user- generated content, and their potential value for dictionaries. *International Handbook of Modern Lexis and Lexicography*, 1-16.

SIL International. (2020). *SIL FieldWorks*. https://software.sil.org/fieldworks/

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary- making*. Cambridge: Cambridge University Press.

Viswanathan, M., Ammerman, A., Eng, E., Garlehner, G., Lohr, K. N., Griffith, D., ... & Whitener, L. (2004). Community-based participatory research: Assessing the evidence: Summary. *AHRQ evidence report summaries*.

6 Annex

Below is the table of the 505 words and their respective Tagalog and English equivalents.

Table 1. The 505 Core Words

CATEGORY	WL#	TAG EQUIV	ENG EQUIV
1. MATERIAL CULTURE			
1.1 Tool	1	daras/piko	adze
	2	palaso/pana	arrow
	3	patalim	blade; arrow
	4	sibat	spear
	5	bendahe	bandage
	6	kudkuran	coconut grater
	7	banig	sleeping mat
	8	unan	pillow
	9	lambat	net
	10	bangka	outrigger canoe
	11	katig	outrigger float of a boat
	12	sagwan	paddle (canoe)
	13	gulong	wheel
1.2 Clothing	14	sinturon	belt
	15	kwintas	necklace (made of beads)
	16	baro	clothes
1.3 House	17	pinto	entrance, gateway
	18	bakod	fence
	19	halamanan	garden
	20	bahay	house
	21	bubong	roof
	22	hagdan	stairs
	23	kamalig	storehouse (food and other products)

1.4 General	24	larawan	image
	25	gayuma	love charm
	26	gamot	medicine; also solution to something
	27	karayom	needle
	28	lubid	rope
	29	sulo	torch, light
2. HUMAN			
2.1 Mental and bodily processes	30	latug	erection
	31	busog	full (after eating)
	32	gutom	to be hungry
	33	gising	awake
	34	panaginip	dream
	35	inaantok	to be drowsy or sleepy
	36	galit	angry
	37	tiis	bear, suffer
	38	bahing	sneeze
	39	maga	swelling
	40	uhaw	thirsty
	41	gising	to wake up
	42	galit	to be angry
	43	hikab	yawn
	44	anak, silang	to bear
	45	dighay	to belch
	46	hinga	to breathe
	47	talik	to copulate (human)
	48	ubo	to cough
	49	dumi, bawas	to defecate
	50	nais, nasa	to desire
	51	patay, panaw	to die/be dead
	52	inom	to drink
	53	lunod (nalunod)	to drown
	54	kain	to eat

	55	takot	fear
	56	tawa	laughter
	57	buhay	life
	58	ibig, mahal	love
	59	kamot	to scratch
	60	tulog	sleep
	61	amoy	scent
	62	limot	to forget
	63	kinig	to hear
	64	tingin, tanaw	to look
	65	kita	to see
	66	alam	to know
	67	dura	to spit
	68	lunok	to swallow
	69	pawis	to sweat
	70	maga	to swell
	71	isip	to think
	72	suka	to vomit
2.2 Body Feature	73	kalbo	bald
	74	kulot	curly hair
	75	payat	thin (human)
2.3 Body part (ext.)	76	sakong, bukong-bukong	ankle
	77	bisig	arm
	78	kili-kili	armpit
	79	likod	back
	80	balbas	beard
	81	tiyan	abdomen
	82	suso	breast
	83	puwit	buttocks
	84	pisngi	cheek
	85	dibdib	chest
	86	bunganga	mouth

87	tainga	ear
88	siko	elbow
89	mata	eye
90	kilay	eyebrow
91	mukha	face
92	taba	fat
93	daliri	finger
94	kuko	fingernail
95	paa	foot
96	noo	forehead
97	uban	gray hair
98	buhok	hair
99	puyo	cowlick
100	kamay	hand
101	ulo	head
102	panga	jaw
103	tuhod	knee
104	binti	leg
105	labi, bibig	lip
106	bibig	mouth
107	kuko	nail
108	batok	nape
109	pusod	navel
110	leeg	neck
111	ilong	nose
112	palad	palm
113	ari ng lalaki	penis
114	balikat	shoulder
115	balat	skin
116	talampakan	sole
117	bayag	testicle
118	hita	thigh
119	daliri sa paa	toe
120	ngipin	tooth (front); all teeth

	121	ari ng babae, pekpek; puki	vagina
2.4 Body part (int.)	122	apdo	bile
	123	dugo	blood
	124	buto	bone
	125	utak	brain
	126	laman-loob	guts
	127	puso	heart
	128	bituka	intestines
	129	atay	liver
	130	baga	lungs
	131	laman	meat (flesh)
	132	tadyang	rib
	133	bungo	skull
	134	tiyan	stomach
	135	lalamunan	throat
2.5 Excretion	136	tutuli	earwax
	137	dume; tae	excrement
	138	utot	flatulence
	139	nana	pus
	140	laway; dura	spittle; saliva
	141	luha	tear (from crying)
	142	ihi	urine
2.6 Health and being	143	bulag	blind
	144	bingi	deaf
	145	pipi	mute
	146	malusog	healthy
	147	sakit	sickness, disease
2.7 General	148	buhay	alive
	149	katawan	body
	150	batang lalaki	boy
	151	bangkay	corpse
	152	batang babae	girl
	153	kati	itch
	154	lalaki	man

	155	tao	person
	156	tinig	voice
	157	babae	woman
3. SOCIAL STRUCTURE			
3.1 Kinship	158	bayaw	brother in law
	159	anak	child
	160	pinsan	cousin
	161	ama	father
	162	byenan	father/mother in law
	163	panganay	first-born
	164	asawa	husband, spouse
	165	bunso	last born
	166	ina, nanay	mother
	167	ulila	orphan
	168	kapatid	sibling
	169	pamangkin	sibling's child
	170	hipag	sister in law
	171	kambal	twins
	172	asawa; may-bahay	wife
3.2 Social Relations	173	pinuno	chief
	174	utang	debt
	175	katulong; alila	servant
	176	alipin	slave
	177	digma; digmaan	war
	178	kasinungalingan	lie
	179	laban	fight; to fight
	180	away	to quarrel
4. COSMOLOGY			
4.1 Environment	181	ulap	cloud
	182	maginaw	cold
	183	araw	sun
	184	hamog	dew
	185	alikabok	dust
	186	lupa	earth; soil
	187	baha	flood

188	bula	foam
189	hamog	fog
190	ginto	gold
191	taog	high tide
192	butas	hole
193	mainit	hot (place)
194	maalingsangan	humid/hot (weather)
195	pulo	island
196	lawa	lake
197	kidlat	lightning
198	buwan	moon
199	putik	mud
200	ilog	river
201	daan	road
202	bato	rock
203	buhangin	sand
204	dagat	sea
205	anino	shadow
206	tabing-dagat	shore
207	himpapawid	sky
208	habagat	southwest wind
209	amihan	northeast wind
210	patpat	stick (of wood)
211	kulog	thunder
212	tubig	water
213	alon	wave
214	hangin	wind
215	gubat	forest
216	ulan	rain
217	ambon	to drizzle
218	laman	content
219	abo	ash
220	uling	charcoal
221	baga	ember

	222	apoy	fire
	223	paso	singe
	224	usok	smoke
	225	sunog	to burn (by itself)
4.2 Physical	226	malaki	big
	227	matalim	sharp
	228	malinis	clean
	229	madumi	dirty
	230	tuyo	dry
	231	mapurol	dull (knife)
	232	matigis	hard
	233	bago	new
	234	luma	old
	235	sira	rotten (as fruit)
	236	bulok	rotten (as log)
	237	talim	sharpness
	238	malambot	soft
	239	matuwid	straight
	240	matangkad	tall
	241	basa	wet
	242	katulad	same
	243	iba	other, different
	244	pangalan/ngalan	name
	245	mapait	bitter
	246	maliwanag	bright
	247	malamig	cold (as of objects)
	248	madilim	dark, dim
	249	mabaho	foul-smelling
	250	mabango	fragrant
	251	sakit	pain
	252	magaspang	rough
	253	maalat	salty
	254	makinis	smooth
	255	maasim	sour
	256	matamis	sweet

257	mainit	warm
258	itim	black
259	pula	red
260	puti	white
261	araw	day
262	umaga	morning
263	lahat	all
264	walo	eight
265	gabi	night
266	malayo	far
267	mabilis	fast
268	kaunti, iilan	few
269	una	first
270	lima	five
271	apat	four
272	puno	full
273	mabigat	heavy
274	ilan	how many
275	magkano	how much
276	huli	last
277	mamaya	later
278	magaan	light
279	mahaba	long
280	maluwag	loose
281	marami	many
282	buwan	month
283	makitid, makipot	narrow
284	malapit	near
285	siyam	nine
286	wala	none
287	ngayon	now
288	madalas, malimit	often
289	minsan	once
290	isa	one

	291	isang daan	one hundred
	292	isang libo	one thousand
	293	ikalawa, pangalawa	second
	294	pito	seven
	295	mababaw	shallow
	296	malalim	deep
	297	maliit	small
	298	anim	six
	299	mahina	weak
	300	mabagal	slow
	301	sampu	ten
	302	makapal	thick
	303	manipis	thin
	304	ikatlo, pangatlo	third
	305	tatlo	three
	306	masikip	tight
	307	ngayon	today
	308	bukas, kinabukasan	tomorrow
	309	dalawa	two
	310	malawak	wide
	311	taon	year
	312	kahapon	yesterday
	313	dito	here
	314	kaliwa	left
	315	gitna	middle
	316	doon	over there
	317	kanan	right
	318	iyon	that
	319	iyan	that
	320	diyan	there
	321	ito	this
4.3 Metaphysical	322	masama	bad
	323	maganda	beautiful
	324	mabuti	good
	325	totoo	TRUE

	326	malakas	strong
	327	pangit	ugly
	328	mahina	weak
	329	mali	wrong
	330	bathala	god
	331	langit	heaven
	332	kaluluwa	soul
5. FLORA, FAUNA, AND FOOD			
5.1 Flora	333	kawayan/buho	bamboo
	334	balat ng kahoy	bark (tree)
	335	buto	bone
	336	sanga	branch
	337	niyog	coconut
	338	ube	edible climbing plant from fleshy root stock
	339	talong	eggplant
	340	bulaklak	flower
	341	luya	ginger
	342	damo	grass
	343	dahon	leaf
	344	lumot	moss
	345	halaman	plant
	346	ugat	root
	347	tubo	sugarcane
	348	tinik (hayop)	bone
	349	punu	tree
	350	puno, katawan	trunk (of tree)
	351	gulay	vegetable
	352	gugo	woody tendril-bearing wine
5.2 Fauna	353	hayop	animal
	354	buwaya	crocodile
	355	aso	dog
	356	usa	deer
	357	itlog	egg

358	balahibo	fur, fine hair, feather
359	palaypay, palikpik	fin
360	palaka	frog
361	hasang	gills
362	pugad	nest
363	pugita	octopus
364	baboy	pig
365	daga	rat
366	hipon	shrimp
367	ahas	snake
368	pusit	squid
369	buntot	tail
370	tinik	fishbone
371	pagong	turtle
372	kalabaw	water buffalo
373	pakpak	wing
374	uod	bulate
375	langgam	ant
376	paru-paro	butterfly
377	ipis	cockroach
378	langaw	fly (small)
379	bangaw	fly (big)
380	kuto	louse
381	lamok	mosquito
382	gagamba	spider
383	anay	termite
384	ibon	bird
385	sisiw	chick
386	manok	chicken
387	uwak	crow
388	paniki	bat
389	igat	eel (freshwater)
390	palos	eel (saltwater)
391	isda	fish

	392	pating	shark
5.3 Food	393	gata	coconut
	394	karne	meat
	395	gatas	milk
	396	asin	salt
	397	alak	wine
6. FUNCTIONAL			
6.1 Functors, question words	398	at	and
	399	sa	non-focused determiner
	400	hindi	no, not, don’t
	401	ano	what
	402	kailan	when
	403	saan, nasaan	where
	404	sino	who
	405	bakit	why
	406	paano	how
6.2 Pronouns	407	siya	he/she
6.3 Expressions	408	ako	I
	409	sila	they
	410	ikaw	you
	411	kami	we (incl)
	412	tayo	we (ex)
	413	kayo	you (pl)
	414	paalam	goodbye
	415	salamat	thank you
	416	walang anuman	welcome
7. VERBS AND MOTION			
7.1 Directional	417	pababa	downward
	418	pataas	upward
	419	diin	to press
	420	tindig	to stand
	421	dala	to bring
	422	dala, buhat	to carry
	423	dating	to come
	424	kaladkad	to drag

	425	hulog, laglag	to fall
	426	lutang	to float
	427	agos	to flow
	428	lipad	to fly
	429	bigay	to give
	430	punta	to go
	431	baba	to go down
	432	pasok	to go inside
	433	labas	to go out
	434	akyat	to go up
	435	sabit	to hang on, hook
	436	talon	to jump
	437	higa	to lie down
	438	bukas	to open
	439	hila	to pull
	440	tulak	to push
	441	lagay	to put
	442	balik	to return
	443	takbo	to run
	444	lubog	to sink
	445	upo	to sit
	446	tayo	to stand
	447	langoy	to swim
	448	lakad	to walk
	449	tuwad	upside down, stooping with the head forward
7.2 Affective	450	dakip	catch, apprehend
	451	halik	kiss
	452	dikdik	pound, well ground
	453	tusok	prick, pierce
	454	banlaw	to rinse
	455	kamot	to scratch
	456	unat	to stretch
	457	hampas	to strike
	458	palo	to beat (strike)

	459	kagat	to bite
	460	baon	to bury
	461	bili	to buy
	462	pili	to choose
	463	linis	to clean
	464	hanap	to find
	465	tama	to hit
	466	hawak	to hold
	467	patay	to kill
	468	laro	to play
	469	bayo, pukpok	to pound
	470	kuskos	to rub
	471	bili, benta	to sell
	472	pakita	to show
	473	piga	to squeeze
	474	saksak	to stab
	475	nakaw	to steal
	476	tuhog	to string
	477	sipsip	to suck
	478	tapon, hagis	to throw
	479	tali	to tie
	480	hugas	to wash
	481	pahid, punas	to wipe
	482	balot	to wrap
7.3 Non-affective	483	tulo	to drip, leak
	484	ihip	to blow (wind)
	485	bilang	to count
	486	sayaw	to dance
	487	awit, kanta	to sing
	488	kinig	to hear
	489	aso	to hunt (game)
7.4 Factive	490	giba	to demolish
	491	tunaw	to melt
	492	kulo	to boil
	493	bali	to break (as stick)

	494	putol	to cut
	495	gawa	to do
	496	tahi	to sew
	497	hati	to split (in half)
	498	habi	to weave
7.5 Communicative	499	tanong	to inquire
	500	tawag	to call
	501	ungol (aso)	to howl (of dogs, wolves)
	502	sabi	to say
	503	sigaw	to shout
	504	salita	to speak
	505	kindat	to wink

A CRITICAL REVIEW OF THE INCLUSION OF IDIOMS IN BILINGUALIZED ENGLISH-CHINESE LEARNER'S DICTIONARIES

Shang Yang
Fudan University, China
yangshang22@foxmail.com

Abstract

Idiomatic expressions to a great extent enrich languages and reflect different cultures. The abundance and variety of English idioms may be due to the historical development of the language. As an essential reference tool for China's English learners, bilingualized English-Chinese learner's dictionaries (BECLDs) record the most frequently used English vocabulary. It seems that the major BECLDs such as *Oxford Advanced Learner's English-Chinese Dictionary* (OALECD) and *Longman Dictionary of Contemporary English (English-English/ English-Chinese bilingualized dictionary)* (LDOCE) tend to cover more and more idioms in their updated versions. However, there are still some improvements that should be taken into consideration as regards their coverage of idioms.

Thus, this paper aims to investigate the inclusion of idioms in the latest editions of the four popular bilingualized learner's dictionaries, namely OALECD, LDOCE, *Macmillan English-Chinese Dictionary for Advanced Learners* (MECDAL), *COBUILD Advanced Learner's English-Chinese Dictionary* (CALECD). The author will critically identify the major problems found therein, which include blurry differentiation between idioms and other fixed expressions, absence of frequently-used idioms, imbalanced inclusion of idioms of related plural nouns, and inadequacy of new idioms, and provide some suggestions for the further revision of the dictionaries mainly adopting a two-pronged approach which are recording more frequently-used idioms and collecting more new idioms.

As Béjoint said, "A dictionary is meant to be consulted" (2001: 19). Idioms, as an essential part of reflecting English cultures that language learners may often encounter, should be given more prominence. It is hoped that this paper could make some contribution to a better inclusion of idioms in future dictionary compilation.

Keywords idioms, inclusion, bilingualized English-Chinese learner's dictionaries

1 Introduction

It is fair to say that idioms enrich languages and reflect different cultures. The abundance and variety of English idioms may be due to the historical development ofthe language. Many of them come from different aspects of the everyday life of the English people such as food and cooking (*as cool as a cucumber, be like chalk and cheese, curry favour*), sports (*beat/ jump the gun, on one's toes, neck and neck*), the Bible (*a land flowing with milk and honey, forbidden fruit, pie in the sky*), literature (*hit the mark, glid the lily, salad days*), fables and myths (*a dog in the manger, bell the cat, the heel of Achilles*)

Idioms are of perennial interest to linguists and lexicographers. A good place to start is to ask what idiom is. Scholars have different criteria to define it. Larson defines an idiom as "a string of words whose meaning is different from the meaning conveyed by the individual words" (1984: 20). Alexander thinks idiom is "multi-word units which have to be learned as a whole, along with associated sociolinguistic, cultural and pragmatics rules of use" (1987: 178). Richards and Schmidt regard an idiom as "an expression which functions as a single unit and whose meaning cannot be worked out from separate parts" (1990: 246). Baker defines idioms as "frozen patterns of language which allow little or no variation in form, and in the case of idioms, often carry meaning which cannot be deduced from their individual components" (1992: 63). Al-kadi defines idioms as being "not literally translatable, as their meanings are unpredictable from the usual meaning of their constituent parts, particularly idioms of socio-cultural, historical, or political backgrounds" (2015: 513).

According to the *Oxford English Dictionary* (OED), idiom is defined as "a form of expression, grammatical construction, phrase, etc., used in a distinctive way in a particular language, dialect, or language variety; *spec.* a group of words established by usage as having a meaning not deducible from the meanings of the individual words". *Merriam-Webster Dictionary* (Merriam-Webster) defines idiom as "an expression in the usage of a language that is peculiar to itself either in having a meaning that cannot be derived from the conjoined meanings of its elements (such as *up in the air* for "undecided") or in its grammatically atypical use of words (such as *give way*)".

The compilation of dictionaries of idioms can be dated back to 19th century and *A Dictionary of English Phrases with Illustrative Sentences* (Kwong Ki Chiu, 1881) and *Dictionary of Idiomatic English Phrases* (James Main Dixon, 1891) are regarded as the most important ones in the history of English lexicography (Gao 2010: 160). The most famous ones nowadays are *Oxford Dictionary of Idioms, Collins COBUILD Dictionary of Idioms, Cambridge International Dictionary of Idioms, Longman Pocket Idioms Dictionary* etc. As an essential part of the dictionary, there is a tendency of covering more and more idioms because "English idioms are very rich and widely used, so they should be included as much as possible" (Lu 2011: 189)

It is unquestionable that new words and new meanings are recorded in the major bilingualized English-Chinese learner's dictionaries, but when it comes to the inclusion of idioms, there are still some improvements could be done. Such problems will be illustrated in detail in the following pages.

Generally speaking, there are three typical dictionaries in China, namely English-Chinese or Chinese-English dictionaries compiled by domestic scholars, original English dictionaries introduced from abroad, and bilingualized English/English-Chinese dictionaries translated from original ones. Bilingualized English-Chinese learner's dictionaries (BECLDs) are one of the most important reference tools for English learners. Lexicographers constantly try to revise them for a better compilation to meet users' needs. *Oxford Advanced Learner's English-Chinese Dictionary (9th Edition)* (OALECD) labels the vocabulary of Oxford 3000 List and Academic Word List. It provides learners with concise and clear definitions and meanings, systematic and comprehensive grammatical information, and offers "Express yourself" notes and Speaking Tutor etc. to further demonstrate the practicality of the language. *Longman Dictionary of Contemporary English Longman Dictionary of Contemporary English (English-English/ English-Chinese bilingualized dictionary) (6th Edition)* (LDOCE) marks Longman Communication 9000 and 3000 most frequent words in spoken and written English, and presents concise explanations, rich examples and collocations, the information of grammar, synonyms, register of words in order to become an effective tool for English learners in reading, writing, and translating. The latest edition of *Macmillan English-Chinese Dictionary for Advanced Learners* (MECDAL) published in 2018, star-rating the frequency of 7500 words in red. It provides the collocation box, innovative usage notes on metaphor and academic writing skills etc. to give learners as much useful information as possible and maximize the consulting and referencing function. *COBUILD Advanced Learner's English-Chinese Dictionary (8th Edition)* (CALECD) is characterized by its definitions and examples which are all from the big Collins Corpus. It better presents new words, new meanings and new usages for users. The full-sentence definitions provide learners a much clearer use of words in the context to guarantee learners to gain the accurate English.

Further to this, dictionary, as “a means of strengthening the language” (Hartmann 1985: 5) and “an aid to foreign-language learning” (Hartmann 1985: 5), plays an essential role of helping language learners to master and apply the target language. Beginners, intermediate language learners, secondary and university teachers all use bilingual dictionaries more frequently than monolingual ones (Hartmann 1983: 46). That is to say, bilingual dictionaries are placed in a very important position by most users.

In the light of the facts outlined above, it is necessary to critically analyze the inclusion of the four major BECLDs and propose some suggestions for a better coverage of idioms in the future bilingual dictionary compilation.

2 Method

This research will present a quantitative study of the inclusion of idioms in the latest editions of four popular bilingualized learner’s dictionaries namely OALECD, LDOCE, MECDAL, and CALECD. The author compares and contrasts the inclusion of idioms in the four bilingualized dictionaries with *Oxford Idioms Dictionary, 2nd edition* (OID) and *An English-Chinese Dictionary of Idioms in Current Use* (ECDICU). The data will be collected mainly from those dictionaries mentioned above to analyze features of the inclusion therein and critically identify the majorproblems found in the sample dictionaries.

3 Result

In general terms, two characteristics of the inclusion of idioms in the four bilingualized English-Chinese learner’s dictionaries can be briefly identified.

Firstly, it is noteworthy that some frequently-used and well-known idioms are recorded in the four BECLDs. Some of the instances are *armed to the teeth, below the belt, cross the Rubicon, down to the wire, go bananas, get the green light, have an ace up one’s sleeve, kick the bucket, meet one’s Waterloo, on the ropes, to laugh up one’s sleeve, with a fine-toothed comb,* etc. And it is fair to say that they are the most familiar ones for English learners in China when they are exposed to English culture. Idioms make daily communication more vivid and convey different emotions. McDevitt (1993) points out that people frequently use idioms in everyday situations. Such widely-used idioms add spice to the English language and guide learners to know the most common ones in daily communication.

Secondly, the four dictionaries place emphasis on providing various labels in regard to style, geographical, and pragmatics labels. For example, style labels include formal, informal, law, computing, literary, spoken etc. Geographical labels cover British English, American English, Australian English, New Zealand English etc. Pragmatics labels consists of approval, disapproval, emphasis etc. Take OALECD for an example, the instances below show the mark with special symbols in the dictionary.

- a bright spark BrE *informal often ironic*
- give sb/ get the bum’s rush *informal especially NAmE*
- off the back of a lorry *BrE informal humorous*
- in the background *computing*
- not have a bar of sth *AustralE NZE informal*
- lay sth bare *formal*

Idioms often cause great difficulties for non-English speaking learners not only because they cannot infer the meaning of idioms from individual words, but also due to the usage scenarios. As Cowie points out that “if, as is now commonly agreed, even advanced students have great difficulty in understanding, and particular difficulty in using high-frequency words, because of the multiple meaning, derivatives, compounds and idioms which they give rise to, then this is the area to which the EFL lexicographer’s principal efforts must be directed” (1983: 136). It would be no exaggeration to say that the accurate understanding and appropriate use of English idioms directly reflect the level of language proficiency and the ability of mastering language. The culture and idiomatic usage affect learners' comprehensive understanding and correct use of idioms. Labels can help English learners deal with the difficulties of idioms and words. As consequence, with the clear labels, learners will know the accurate meanings and the actual use in spoken and written contexts. Through these labels, leaners will know unique cultural connotations of native English-speaking countries across the spectrum of historical development, religious belief, custom and habits, fables and mythologies, literary works and so forth more clearly and avoid misusing idioms.

4 Analysis and Discussion

Although the idioms included in the four big dictionaries cover a wide range of aspects of English-speaking countries mainly the UK, the USA and Australia, it is never without some drawbacks. Let us unpack each problem in detail for a better analysis and discussion of the inclusion of idioms in the four BECLDs, namely blurry differentiation between idioms and other fixed expressions or even colloquial expressions, absence of frequently used idioms, imbalanced inclusion of idioms of related plural nouns, and inadequacy of new idioms.

4.1 Blurry Differentiation Between Idioms and Other Fixed Expressions

It seems that lexicographers haven't reached an agreement of the inclusion principle of idioms no matter in those idiom dictionaries before 1949 or in the latest ones. And the loose standard of idiom inclusion principle in dictionaries seems have little improvement.

Take OALECD for an example, some idioms are obviously not idiomatic expressions, but the dictionary puts them in the "idiom section". According to the definition of idiom in *OED* and *Merriam-Webster* mentioned above, the meaning of an idiom cannot be deducible from the meaning of its individual elements. At least, colloquial or spoken expressions and phrase or sentence patterns should not be marked as "idiom". Such loose standard of inclusion policy to some extent will mislead learners. And it is doubtful whether such an inclusion principle would be blurry and unreasonable. Table 1 shows some examples.

Table 1 OALECD examples

Entries	Idioms
hear	have you heard the one about…? hear tell (of sth) *old-fashioned or formal* I've heard it all before *informal* (do) you hear (me)?
how	How about? How do you do? How can/ could you How's that
just	It is just as well (that…) just about *informal* just a minute/moment/second just like that just now just then
know	I don't know how, why, etc. *informal* I know *informal* let it be known/ make it known that… *informal* you know something/ what? you know who/ what?
question	good question! just/merely/only a question of (sth/doing sth)
yeah	oh yeah? year right

4.2 Absence of Frequently-Used Idioms

An important point to note is the absence of frequently used idioms. Although the four BECLDs record many common idioms, some are failure to be found. Table 2 below presents the comparison of random examples between BECLDs and OID.

Table 2 Comparison with OID

OID	OALEC D	LDOCE	MECDAL	CALEC D
fiddle while Rome burns	-	-	-	-
fire in the belly	-	+	-	-
handle with kid gloves	+	+	-	-
move mountains	-	+	+	-
squeal like a stuck pig	-	-	-	
the tail wags the dog	-	+	+	-
three sheets to the wind	-	-	+	-

NB: +marks the idioms already been included; - marks non-included

Idioms are different from words, phrases and collocations that their meanings can be inferred through the context. However, the meaning of idioms cannot be deducible from the individual words. This is to some extent the most difficult part for English learners to understand idioms. Thus, learners have to consult dictionaries to find out the meaning. From the table above which can be seen, the absence of widely-used idioms may be far from enough to meet learners' needs especially for advanced leaners. In the meantime, it is not surprising that such limitation would affect the retrieval results of dictionaries.

4.3 Imbalanced Inclusion of Idioms of Related Plural Nouns

After unpacking idioms in the four dictionaries, it seems that they pay more attention to the idioms of singular nouns, but more or less ignore the idioms of related plural nouns. This is justified by the example of the "book" entry. Table 3 displays the inclusion of four BECLDs recording idioms of "books".

Table 3 Examples of "books"

BECLDs	"books"
OALECD	be in sb's good/bad books *informal* be on sb's books cook the books/fiddle the books
LDOCE	cook the books on the books be in sb's good/bad books *informal*
MECDA	do the books, in sb's bad/ good books
CALECD	cook the books

From the table above, we can see that four BECLDs mainly include *be in sb's good/bad books, be on sb's books, cook the books,* and *do the books,* but those idioms such as h*it the books/ crack the books, off the books, one for the books/ one for the book* are failed to be recorded in the dictionaries. They are as frequently-used as the included ones in the daily life.

Another example is particularly the case which is "dogs" as Table 4 demonstrated. Four dictionaries all fail to record *throw sb. to the dogs* therein which is a common idiom as well. More examples can be instantiated such as "boots" and "words" etc. *Die with one's boots on, hang up one's boots,* and *quake/shake/quiver in one's boots* cannot be found in the four BECLDs.

Table 4 Examples of "dogs"

BECLDs	"dogs"
OALECD	go to the dogs *NAmE* also go to hell in a handbasket *informal* be raining cats and dogs
LDOCE	be going to the dogs *informal* the dogs *BrE informal* let sleeping dogs lie
MECDA	the dogs *BrE informal* going to the dogs *informal* be raining cats and dogs let sleeping dogs lie
CALECD	is going to the dogs

4.4 Inadequacy of New Idioms

The number of new idioms is somewhat far from enough although some can be found in the latest edition. Some new idioms keep emerging but they have not been collected in the BECLDs yet. ECDICU is used to compare with the four BECLDs in regard to the collection of new idioms. Table 5 shows some examples.

Table 5 Comparison with ECDICU

ECDICU	OALECD	LDOCE	MECDAL	CALECD
shoot down in flames	-	-	+	-
play both ends against the middle	-	-	+	-
see/view the glass as half-empty/ half full	-	+	-	-
feel one's legs	-	-	-	-
gum up the works	-	-	-	-
let nature take its course	-	+	+	-
peaches and cream (an enjoyable and pleasant experience)	-	Used to describe skin that is an attractive pink colour	Skin is smooth, pale and slightly pink	Clear, smooth, pare skin

NB: + marks the idioms included; - marks non-included; meaning of idioms marked in the table demonstrates that it hasn't been included in four BECLDs.

From the table above, we can see the reality is that some new idiomatic expressions are failure to collect in the four big dictionaries. Besides MECDAL, the other three dictionaries all mention the feature of extensive collection of new words in their prefaces. Lexicographers' constantly pay attention to new words and new meanings when they revise dictionaries. It is worth mentioning that idioms should not be an exception because they reflect the authentic British and American culture, and they are a microcosm of the culture. With the change and development of language, many new idioms continually keep appearing in people's daily life. Such phenomenon is also a challenge to the inclusion of idioms. They can help learners to understand many idiomatic expressions that are really used in the process of consulting dictionaries. However, some new idiomatic expressions that people use in daily life are absent in dictionaries. In consequence, it may bring inconvenience to users and make the information limited. In other words, limited inclusion of new idioms may hinder English learners from mastering new idioms.

5 Suggestions for Better Coverage

After deeply and systematically analyzing the coverage of idioms and discussing the major problems in regard to the inclusion of BECLDs, what brings lexicographers to think about is that what they should take into consideration when recording idioms? From my perspective, for the future revision of bilingualized English-Chinese learner's dictionaries, the editors should adopt a two-pronged approach that will be spelled out below.

5.1 More Frequently-Used Idioms Should Be Recorded

As discussed above, lacking frequently-used idioms is one of the main problems when it comes to the inclusion matter. Thus, more commonly-used idioms should be welcomed into the dictionaries.

For language learners' part, it is necessary for them to accumulate more knowledge and keep pace with the emerging of new expressions. As McDavid (1979: 19-20) said in his paper, one of the functions of a dictionary is to get learners exposed to the target language. Some years later, Hartmann (1985: 5) proposed seven language functions, and two of them are "the dictionary as a means of strengthening the language" and "the dictionary as an aid to foreign-language learning". Their points of view arrive at essentially identical conclusions that dictionaries help learners master and apply a language. It is essential to place emphasis on the knowledge accumulation especially from the cultural perspective as learning a language means learning the culture which is behind the authentic expressions. Idioms, to some extent, is an integral part of acquainting a language when consulting dictionaries. Understanding and using idioms could be regarded as an indicator of learners' language proficiency. Therefore, as an indispensable part of dictionaries, it goes without saying that English idioms should be recorded more thoroughly. Recording some more widely-used idioms can help English learners to better understand and master them, and effectively use them in communication. In similar vein, it helps learners to clarify the meaning of them so as to be able to use them correctly after consulting them in the dictionaries.

5.2 More New Idioms Should Be Included

New idioms should be placed emphasis as language is dynamic and reflects the culture of English-speaking countries. It is manifested in the fact that the inclusion of new idioms now is rather pale after analyzing some examples from four BECLDs above. With more and more new words and phrases recorded in the updated edition of the dictionaries, the coverage of idioms should be of paramount importance as well. To address the major problems mentioned above, the top priority is to give idioms the same importance to keep pace with the inclusion of other new entries. One feasible way for editors is to refer to the World Wide Web or news archives to check the frequency and currency of new idioms. In this regard, some new and popular idioms with high frequency should be included in the future dictionaries. As Hornby and Parnwell (1972: 135) said that a dictionary can not only tell users the spelling of words but also expand their language knowledge that any textbook cannot compare with.

From the analysis in part three, it is not unusual that new words are always put in an important place in the revision of the dictionary, while new idioms seem not have received the same attention. Some new idioms that have frequently used before the publication of the latest editions have not been included, which will inevitably affect the retrieval rate of the dictionary. The advanced dictionaries are mainly made for advanced English learners. They are supposed to cover more new and updated information. If the coverage of new idioms can be much improved, it would be more conducible and helpful to advanced learners.

In a word, dictionary as a tool of recording language, although it is impossible to record all idioms, we can expand the scale of inclusion. Through this way, it could expand the amount of information of the dictionary, enhance its reference function, and improve the search rate. To achieve this goal, it requires lexicographers to pay more attention to the new idioms appearing in recent years for a better revision in the next edition.

6 Conclusion

This paper begins from investigating the coverage of idioms in the latest edition of the four big bilingualized learner's dictionaries, namely OALECD, LDOCE, MECDAL, and CALECD. In the following parts, the paper has critically evaluated the problems from detailed aspects, made some comparison with ECDICU and OID, and proposed some suggestions based on the previous discussion for the future dictionary compilation. The results show that there are not only general problems the four BECLDs have, but also some individual problems in each BECLD including blurry differentiation between idioms and other fixed expressions, absence of frequently used idioms, imbalanced inclusion of idioms of related plural nouns, and inadequacy of new idioms. Instantiations are found to demonstrate such problems.

The four dictionaries that were discussed above are chiefly for advanced English learners. From my perspective, they are supposed to attach greater importance to the inclusion of idioms since they are an indispensable part of the dictionary and a reflection to the culture of English-speaking countries. With the ebb and flow of some idioms, it is compilers' work to judge whether an idiom should be collected in a dictionary. Further to this, the frequently-used and new idioms should be given a repeated consideration in the future revision. What is more, with help of corpus observing the development of idioms, it would be more reliable to decide the spectrum of idioms.

7 References

Alexander, R. (1987). Problems in Understanding and Teaching Idiomaticity in English. *Anglistik and Eneglichunterricht,* 32(2), 105-122.

Al-Kadi, A. (2015). Towards Idiomatic Competence of Yemeni EFL Undergraduates. *Journal of Language Teaching and Research,* 6(3), 513-523. https://doi.org/10.17507/jltr.0603.06

Baker, M. (1992). *In Other Words: A Course Book on Translation.* London: Routledge. Béjoint, H. (2002). *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

COBUILD Advanced Learner's English-Chinese Dictionary (8th ed.). (2017). HarperCollins Publishers Limited and Foreign Language Teaching and Research Press.

Cowie, A. P. (1983). English dictionaries for the foreign learner. In Hartmann (Ed.), *Lexicography: Principles and Practice* (pp. 135-143). London: Academic Press. Gao, Y. (2010). English-Chinese Idiom Dictionaries before 1949. *Lexicographical Studies* (5), 159-172.

Gao, Y. (2014). *An English-Chinese Dictionary of Idioms in Current Use.* Shanghai: Shanghai Translation Publishing House.

Hartmann, R. R. K. (1983). *Lexicography: Principles and Practice.* London; Orlando: Academic Press.

Hartmann, R. R. K. (1985). *Dictionaries of English: the user's perspective*. manuscript.

Hornby, A.S., & Parnwell, E. C. (1972). *Progressive English Dictionary*. Oxford: Oxford University Press.

Hornby, A. S. (2018). *Oxford Advanced Learner's English-Chinese Dictionary* (9th ed.). Oxford: Oxford University Press. Beijing: The Commercial Press.

Larson, M. (1984). *Meaning-Based Translation: A Guide to Cross Language Equivalence.* Lanham, New York and London: University Press of America. *Longman Dictionary of Contemporary English, sixth Edition*. (2019). Pearson Education Limited. Beijing: Foreign Language Teaching and Research Press.

Lu, G., & Wang, F. (2011). *A Study on the Compilation Characteristics of Big-Sized Bilingualized Dictionaries -- A Case Study of The English-Chinese Dictionary Compilation.* Shanghai: Shanghai Translation Publishing House.

McDavid, R. I., Jr. (1979). The Social Role of the Dictionary. In Congleton et al. (Eds.) (pp. 17-28).

McDevitt, E. (1993). *What Does That Mean? An Introduction to American Idioms.* Department of Education, Washington, DC.

Macmillan English-Chinese Dictionary for Advanced Learners. (2018). London: Macmillan Publishers Limited. Beijing: Foreign Language Teaching and Research Press.

Merriam-Webster Dictionary. Online at <https://www.merriam-webster.com/>.

Oxford English Dictionary. Online at <https://www.oed.com/>.

Oxford University Press. (2013). *Oxford Idioms Dictionary* (2nd ed.). Beijing: Foreign Language Teaching and Research Press. Oxford: Oxford University Press.

Richards, J., & Schmidt, R. (1990). *Longman Dictionary of Language Teaching and Applied Linguistics.* London: Longman.

RESEARCH INTO THE SITUATIONS AND PROBLEMS OF COMPILATION AND PUBLICATION OF DIGITALIZED DICTIONARIES IN THE AGE OF MEDIA CONVERGENCE

Shuhang Tang
Guangdong University of Foreign Studies, China
TSHJulie@163.com

Abstract

The media convergence provides a major opportunity for the digital transformation of dictionaries. Current studies on the digitalized dictionaries mainly focus on the multimedia or multimodal ones, frequently accompanied by obscuring the functions between multimedia and multimodalities for lexicographical compilation and publication in the age of media convergence. It would be fair to say that the digitalized dictionaries remain the primary phase. Since the current compilation situations reveal such problems as mixing the concept of multimedia and multimodalities up, lack of multimodal-lexicographical text based on media convergence, inconsistency between the unit and approach of the organization for lexicographical text and precise service for user groups. While in the process of publication the digitalized dictionaries show the short of the ununified compilation and publication platform, new media access to the lexicographical text as well as new profit model throughout the investigation in this paper.

Keywords digitalized dictionaries, compilation and publication, media convergence, situations and problems

1 Introduction

The concept of "media convergence" has seen the history of 38 years since it was put forward in 1983.i[1] The essence of media convergence is "convergence". It aims at making an accommodation among various media of communication by means of technologies of Internet Plus and artificial intelligent to allow textual information communicate across the borderlines of media. To be specific, media convergence can be called media of whole journey or omnimedia since it benefits anybody who makes use of Internet terminals to read, use and exchange information (media of entire personnel) through multimedia (holographic media) and multimodalities (media of full effect) (Veglis et al. 2016; Zhang 2019). President Xi Jinping in China raised the following issue when he made an on-the-spot investigation for *People's Daily* in the 25th January, 2019: researching and developing omnimedia has already been an urgent research project right now. He laid an emphasis on the convert from "promoting the major task of media convergence development" to "promoting media convergence to develop in deep way"ii[2]. The Lexicographical Society of China (LSOC) makes such a quick response that seminars such as "media convergence and dictionaries" in March and "dictionaries based on media convergence" in April have been launched successively. It is worth mentioning that other national lexicographical seminars held in Shanghai and Guangzhou put dictionaries based on media convergence for the first subject under discussion as well. It is natural to draw a conclusion that media convergence is of great significance for the digital transformation of dictionaries. Therefore, this paper attempts to discover the current problems existed in the transformation of traditional dictionaries through the investigation of both compilation and publication of digitalized dictionaries so

[1] i Ithiel de Sola Pool in Massachusetts Institute of Technology proposed "the convergence of modes" in Technologies of Freedom in 1983, which formed a preliminary notion of "media convergence".

[2] ii To refer to the relevant report in the news net of Communist Part of China http://cpc.people.com.cn/n1/2019/0125/c64094-30590946.html

as to lay a foundation for the renovation and development of dictionaries based on media convergence.

2 Current Situations and Problems of Compilation for Digitalized Dictionaries

The primary types of the digitalized dictionaries include hand-held e-dictionaries, CD-ROM dictionaries, online dictionaries, exe install program and APP dictionaries. In 1990s, the authoritative mainstream dictionaries touched Internet for the first time, such as *Oxford English Dictionary* (OED) and *Encyclopedia Americana* published CD-ROM dictionaries in 1992 and 1995 respectively. Thereafter, the series of Oxford and Longman promote CD-ROM dictionaries and Internet ones, and our country also promote CD-ROM version of *The Chinese Dictionary* and Hongen online, which initiates the course of electronic transformation of paper dictionaries. In 2012, *Encyclopedia Britannica* with 244 years of history turned to full-scale digital publication at the same time Macmillan made a dictionary of online-version-only announcement for the first time. Although international dictionary academy has not proposed the concept of "dictionary based on media convergence" yet, they develop such versions as CD-ROM, online, mobile phone and desktop on the basis of CD-ROM and Internet media throughout such mainstream dictionaries as Oxford, Longman, Collins. The digitalized dictionary has been an important form of publication in international dictionary publishers, and then lexicographical texts have been integrated the elements of multimedia and multimodalities such as picture, audio frequency, video frequency and man-machine interaction (Dai & Xu 2014; Zhang 2019). *Chinese Encyclopedia* (3rd edition), China's 12th Five-Year Plan of publication in 2011 for major dictionaries, has been planned to promote modified version on Internet. Other authoritative learners' dictionaries such as new *Modern Chinese Dictionary* and *Xinhua Dictionary* as well as reference dictionaries such as *Ci Hai* and *The Chinese Dictionary* begin to publish successive APP version and electronic one since 2019. All of them have been greatly welcomed by the public due to their utility, convenience and innovation. Nevertheless, when applying such digital elements as multimedia and multimodalities for lexicographical texts, these dictionaries are not enough.

Therefore, the digital Internet versions and APP ones of newly published international mainstream dictionaries for learners of English mostly used among domestic user groups are to be selected in this paper to investigate. The content of investigation includes such interface setups as compilation principle, stylistic rule and layout, lexicographical text, multimodal interaction. These dictionaries are major five dictionaries for learners of English (Big Five), merely, APP version of *Oxford Advanced Learner's English-Chinese Dictionary* (9th edition) (The Commercial Press, 2019), ISO version of *Longman Dictionary of Contemporary English* (5th edition) (Foreign Language Teaching and Research Press, 2014), mobile version of *Cambridge Advanced Learner's Dictionary* (4th edition) (Foreign Language Teaching and Research Press, 2010), APP version of *Collins COBUILD Advanced Learner's English-Chinese Dictionary* (8th edition) (Foreign Language Teaching and Research Press, 2017) and online version of *Macmillan English-Chinese Dictionary* (2nd edition) (Foreign Language Teaching and Research Press, 2018). The results are shown in table 1.

Table 1 Compilation Interface Setups of Big Five Digitalized Dictionaries

names of dictionaries modules	APP version of *Oxford Advanced Learner's English-Chinese Dictionary* (9th edition)	ISO version of *Longman Dictionary of Contemporary English* (5th edition)	mobile version of *Cambridge Advanced Learner's Dictionary* (4th edition)	APP version of *Collins COBUILD Advanced Learner's English-Chinese Dictionary* (8th edition)	online version of *Macmillan English-Chinese Dictionary* (2nd edition)
text	definition of definition style and antithesis one; English-Chinese exemplifier; topic collocation column; usage note and sense disambiguation	phrase definition; bilingual example; etymology; word frequency; sense disambiguation	principle of "one entry, one core meaning"; word frequency; idiom index; column of common error and usage note for learners; sense disambiguation	sentence definition; bilingual example; word frequency; lexical chain; usage note and sense disambiguation	four major initiative columns: lexical collocation， word extension academic writing and metaphor ； word frequency; sense menu; usage note and language tip

graph	graphic explanation for lexicon	natural classical color collocated with eye-protecting color; graphic explanation for word and phrase	illustration	full color; differential graphic definition	red basic lexicon; red asterisk labeling word frequency; illustration and diagram; graphic explanation for word
audio frequency	Oxford authentic pronunciation depot; human pronunciation for lexicographical dialogue	angle-free role voice; pronunciations of British style and American one; illustrative sentence recording	human pronunciations of British style and American one; pronunciation for choice of word	human pronunciations of British style and American one	human pronunciations of British style and American one; pronunciation for choice of word
video frequency	cloud service	hyperlink	no	split screen view	action illustration
man-machine interaction	personalized query and search as well as learning function (e.g. iwriter as wording suggestion function for writing)	real-time progressive search; fuzzy search; functions of lexicon activator (e.g. synonym comparison and activator) and virtual reality (e.g. manual trigger and visual space)	offline query; functions of vocabulary notebook and information collection; available 3D Touch gesture	powerful search function (e.g. functions of word searching jumping and hyper translation); available 3D Touch and Peek gesture; available iCloud syncing bookmark	no
man-man interaction	no	e-mail correspondence (user-user; compiler-user)	no	no	no

It can be revealed the following four problems through the investigation of Big Five.

(1) Mixing the concept of multimedia and multimodalities up. It can be seen from the above table that whatever versions of APP, ISO, mobile and online, the paper text of these dictionaries realizes convergence on the digital interface of spatial structure, functions of multimodalities and intelligent technology to some degree. It is such module functions as Graph, picture, pronunciation, sound recording, pronunciation comparison, man-machine conversation and virtual reality that counts. However, we notice that audio frequency, video frequency and something like that are just media of lexicographical text communication which includes various modalities. For instance, audio frequency contains simulation modality and acoustic one; video frequency contains such modalities as dynamic picture, structure graph, dynamic voice and text (Kress & van Leeuwen 2006; Lew 2010). Unfortunately, there are still the phenomenon of confusing "multimedia" and "multimodalities" among Big Five, taking the functions of picture and pronunciation as multimodalities. For example, mobile version of *Cambridge Advanced Learner's Dictionary* (4th edition) is stand-alone Internet dictionary to be use offline by means of Internet download or leading-in lexicographical data packet. It only offers illustration without the modalities under video frequency particularly. While online version of *Macmillan English-Chinese Dictionary* (2nd edition) lacks such multimodalities as man-machine interaction and man-man dialogue. The most important question is that all of them neither convergent the modalities among media, nor put modality and written script in the same place to present the meaning of entry. Mixing the concept of multimedia and multimodalities up means digital transformation of dictionaries just electronic transformation of paper version.

(2) Lack of multimodal-lexicographical texts based on media convergence. Big Five present in the same way in the aspect of interface structure. After transplanting paper content to Internet or electronic device, they make use of picture and pronunciation to enhance query and representation, to add such new

forms matching these contents as video frequency and animation. For instance, the function of information collection in mobile version of *Cambridge Advanced Learner's Dictionary* (4th edition) provides rich merch of knowledge service. Except for online version of *Macmillan English-Chinese Dictionary* (2nd edition), other four dictionaries have the function of man-machine interaction. In particular, ISO version of *Longman Dictionary of Contemporary English* (5th edition) adds the module of man-man interaction, which bridging the communication between compiler and user. But none of them demonstrate text characteristics of multimodalities based on media convergence in which character is still one of the mostly used definition modalities in formal lexicographical text. Big Five have already put picture modality and acoustic one into it, ISO version of *Longman Dictionary of Contemporary English* (5th edition) and APP version of *Collins COBUILD Advanced Learner's English-Chinese Dictionary* (8th edition) supply functions of virtual reality and finger trigger, modality of sense of touch. Even under this circumstance, other modalities do not join in the definition and connotation of entry in direct way, just playing a role of additional information and decoration content. In brief, these unique modalities have not been mixed together to construct the formal text of modalities for lemma.

(3) Inconsistent organization unit and pattern for lexicographical text. Basically, Big Five are compiled in dictionary unit other than in entry or entry meta-data one. Thus, the definition and connotation of entry are incomplete. To be specific, most of these various linguistic property, register, professional property and their degree labeling for lemma have not been made evident annotation and hierarchy yet. Especially such sources and materials as picture, audio frequency and video frequency involving image representation for lexeme are more than anything else. Only APP version of Collins COBUILD Advanced Learner's English-Chinese Dictionary (8th edition) adopts differential graphic definition and online version of *Macmillan English-Chinese Dictionary* (2nd edition) employs red asterisk to mark word frequency. Due to informal and inconsistent digital coding, these dictionaries compile without database giving concrete label to be put in the fixed place by means of spatial organization form. So, users cannot set the representation structure up in the aspects of query items, knowledge distribution and lexicographical text and other various modality information on the interface themselves.

(4) Short of exact service for user groups. Apart from APP version of *Oxford Advanced Learner's English-Chinese Dictionary* (9th edition) and online version of *Macmillan English- Chinese Dictionary* (2nd edition), other three dictionaries are capable of accomplishing man- machine conversation and virtual reality. When it comes to user discussion and interaction, only ISO version of *Longman Dictionary of Contemporary English* (5th edition) adopts e-mail to make the indirect dialogues between compilers and users as well as users and users come true. Nevertheless, the communication pattern of media convergence appears extreme subdivision and fragmentation, requiring the unit of entry or entry meta-data to service users at various levels in exactly systematic and fragmented way. The differences of both dictionary users and their performances cannot deal with the change of users' reference habits and new reference demands through the direct interaction between compilers and users as well as users and users in Big Five. If users cannot make and convey our dynamic impressions towards dictionary use or opinions and suggestions towards dictionary compilation timely and effectively, they cannot set exact scope and method up, even customize the individual dictionary on their own according to query preference.

3 Current Situations and Problems of Publication for Digitalized Dictionaries

The increase of our digitalized dictionaries is much more considerable. According to Zhixin Zhou and Lijing Bai (2014), the digitalized dictionaries exceed their paper version for the first time in 2012. The data of "research report of online dictionaries development in 2012" announced by iResearch reveals that the penetration rates of Internet word searching and translation service reach 73.7%, and the market penetration rates of online translation website, online dictionary and download dictionary software all account for more than 50% in 2012. Moreover, the total downloads of APP version of dictionary on mobile phone are over 1 billion. The production of traditional dictionaries has decreased sustainably since 2013. According to the latest data of Investigation Report of Book Publication Market in China in 2019, the total publication of traditional dictionaries descends 3.1% compared with 2018. Meanwhile, the publication of digitalized dictionaries increases more than entire publication industry. Sample statistics shows that there are at least 300 Internet dictionaries or APP ones in our country, referring to such types of dictionary as monolingual, bilingual, comprehensive, specialist and synonym, which is consistent with the types of traditional dictionary. Then it can be summarized the following main three problems through the investigation for their current situations.

3.1 Inconsistent Platform of Compilation and Publication

LSOC established professional committee on modern technology of dictionary compilation in 2001. With regard to aiming at promoting traditional dictionary develop towards dictionary based on media convergence, dictionary publishers and research centers also did a lot of work to explore it. The dictionary corpus and compilation system launched by The Commercial Press makes the whole process from storage to online publishing for lexicographical compilation and publication storage automatic (Liu 2007); the compilation system for bilingual dictionary initiated by Shanghai Foreign Language Education Press is an integrated digital platform, making lexicographical compilation full paper-free so as to bring about real digitalization (Zhuang 2013); the dictionary generation system developed by Center for Lexicographical Studies in Guangdong University of Foreign Studies accomplishes the automatic generation of dictionary on the basis of database (Zhang & Liu 2007). However, these current lexicographical electronic data scattered wide over the country with unique purpose. Their standards and formats differ that hardly form joint force within the same platform to share source.

Besides, there are lexicographical sources and small-scale linguistic data extracted from various media such as book, newspaper, broadcast, television, Internet and social media in small-medium publishing constitutions. The text format and coding method of their compilation and publication are pretty much more of chaos, which brings an extreme challenge for making lexicographical meta-data of multimodalities access to universal media edit and publication.

3.2 Urgent Need for New Media for Lexicographical Text

Dictionary users prefer to labeling usage demand, reference for fragmented information with networking approach. These new changes of using habits make online reference and mobile terminal usual state for all dictionary users (Zhang 2021). Paper dictionaries launch Internet version, online one sequentially, merging multimedia sources and covering various terminals to meet users' demand for query at any timey and any place. Especially large-scale dictionary publishers have a tendency for digitalization and new media, their media of communication applied for digitalized dictionary lead the whole publishing industry that attracts wide attention. The concrete media of communication as shown in table 2.

Table 2 Convergences between Lexicographical Texts and Media of Communication

names of dictionaries media of communication	Internet version of *Ci Hai* (7th edition)	Internet version of *Chinese Encyclopedia* (3rd edition)	Internet version of *The Chinese Dictionary* (2nd edition)	versions of Internet and USB flash disk of *Ci Yuan* (3rd edition)
media of publication	Internet communication WeChat social network	Internet communication social network	Internet communication social network	Internet communication social network simple information polymerization

media of logic	picture audio frequency video frequency amination 3D dynamic model virtual reality (VR) augmented reality (AR) interaction between paper and Internet (link of quick response code)	picture audio frequency video frequency amination anime three-dimensional model interaction between paper and Internet (link of QR code)	picture audio frequency video frequency	picture audio frequency video frequency interaction between paper and Internet (link of QR code)
media of physics	computer terminal terminal of mobile phone electronic reading equipment	computer terminal terminal of mobile phone	computer terminal terminal of mobile phone	computer terminal USB flash disk media player

The first thing to apply media convergence for dictionary is to figure out media of communication involved in lexicographical text. Media of publication among media convergence contains different methods to release information content, which is frequently based on Internet. It can be seen that the Internet versions of these dictionaries are all published without Internet communication and social network. For example, WeChat official account in Internet version of *Ci Hai* (7th edition) exactly demonstrates an active move response to the change of users' reference habits. But these dictionaries do not touch such media of publication as e-mail, short message, microblog, broadcast. In addition, media of communication transmit a variety of media of logic. In spite of such basic media of logic as audio frequency, video frequency, amination and spatial model, there are modalities of sense of sight and touch to be treated in communication under VR and AR. Contrary to overseas countries, other dictionaries, except Internet version of *The Chinese Dictionary* (2nd edition), all make use of link of QR code that is widely used among domestic users to accomplish the interaction between paper media and Internet one, which having an effective impact on compatible dictionary publication and usage between paper version and digital one. The available terminals of these media are primarily computer (including personal computer, tablet personal computer), mobile phone, electronic reading equipment, media player. It is worthy of note that USB flash disk version of Ci Yuan (3rd edition) can be available for USB flash disk. Last but not least, other mobile reading terminals such as electronic organizer, electronic dress and wear items need further developing and converging.

3.3 Further Innovation for Profit Model

The compilation and publication of digitalized dictionary require plenty of capitalized cost to support for a long time in the age of media convergence. Although there are the projects at the level of country and education to help invest and construct, further operation, maintenance, iteration and update need certain marketing or profit revenue to sustain the scale increase of dictionaries based on media convergence for dictionary publishers. To take the most part of influential newly- published digitalized language dictionaries in our country for example, we make sample survey for the current situations of their profit models. The following table 3 shows the result.

Table 3 Sample Survey for Profit Models of Digitalized Language Dictionaries

names of dictionaries / profit conditions	APP version of Modern Chinese Dictionary (7th edition)	APP version of Xinhua Dictionary (12th edition)	Internet version of Ci Hai (7th edition)	electronic version of The Chinese Dictionary (7th edition)
publicity highlights	intelligent dictionary helper; providing fast and convenient functions of query and learning	QR code scanning for reading books, writing strokes of a Chinese character and listening pronunciations	knowledge map; more 3D dynamic modules; available on the webpage of computer, application of mobile phone and official account of WeChat	full offline usage; photograph scanning word searching; definition network query; new words syncing learning, collection and review
profit models	half free, half fee (limit for 2 words per day; 0.1 RMB purchase in advance, 40 RMB for all functions of the dictionary within the fixed time, afterwards 98 RMB)	half free, half fee (limit for 2 words per day; registering and downloading extension packet remove constraint by paying 40 RMB)	earlier free, later fee (limit for 6 times per day; purchasing paper version of the dictionary access to 5-year use of Internet version, purchasing for 98 per year)	full fee (annual charge)
comments of users	"supporting paying for wisdom"; "The price can be more considerable"; "APP version of the dictionary is supposed to be the additional service for purchasing paper one"	"a design of taking users' experience as the center"; "The price is about that of pork per 0.5 kilogram"	"The price is too high"	"If every word is linked to the scanning page of the paper version of books, it will be perfect"; "Do not charge annually, one-off charge is the best"

Due to paper versions of the authoritative dictionaries having already taken up most of domestic market shares, the marketing of these dictionaries themselves occupies the incomes of dictionary publishers. therefore, it can be observed from the above table that all digital versions of these dictionaries adopt the profit models of fee or half fee, selling more expensive. More specifically, APP version of *Modern Chinese Dictionary* (7th edition), it sells 98 RMB, roughly the same as its paper version. Even if APP version of *Xinhua Dictionary* (12th edition) only charges 40 RMB, it also be joked to be "The price is about that of pork per 0.5 kilogram" by users. While both Internet version of *Ci Hai* (7th edition) and electronic version of *The Chinese Dictionary* charge annually, with the former sells 98 RMB per year. In spite of content update of every year, the comments of users (excluding the improvement of dictionary functions) show that many of users regard their prices to be too high, annual charge being not so much as it is one-off trade. What's more, these dictionaries sell login account or block functions of dictionaries to obtain profit. This will be inevitably extruded by those long-free dictionaries with low quality during a period of time, weakening capital source of dictionary compilation.

A great amount of digitalized dictionaries adopt free or half free profit models since most users are inclined to use free dictionaries. And this is exactly the difficult problem dictionary publishers encounter right now (Kilgarriff 2006; Chen 2010; Lv 2020). As a matter of fact, some dictionaries in Chinese Taiwan, and other excellent dictionaries overseas publish paper version as well as Internet one of most dictionaries for free, fewer for fee. Foreign authoritative dictionary publishers earn by means of publicity, click rate or advertisement for their paper versions, and thus the earnings of dictionaries for fee increase a lot. One of the most representatives is OED that is successful in both aspects of fee and profit for Internet version. Yuming Li, president of LSOC, suggested that we ought to exert such an advantage of our own country that our country buy these good dictionaries to free for everyone on the Internet in New Era Seminar on Lexicography and Dictionary Development in 2019.

In fact, new profit channels are called upon for digitalized dictionaries, especially small-medium digitalized ones, depending on the cooperation with a common effort between dictionary publishers and technology suppliers (Internet, linguistic technology, electronic communication technology) under the background of media convergence (Leminen et al. 2016). These technology suppliers include information technology companies, cloud platforms, search companies, terminal manufacturers or operators of digitalized dictionaries as well. They create vast and diverse operating models based on pretty incomes among industrial market, which provides new opportunity of income distribution for dictionary publishers. The giant partnership between compiling superiority of professional lexicographical talent and communication benefit of technology platform initiates publication model shoulder to shoulder of both digital version and paper one. Therefore, digital version is plentiful and comprehensive at the same time paper version is authoritative and standard.

4 Conclusion

The media convergence provides a major opportunity for the digital transformation of dictionaries. Meanwhile, the compilation and publication of digitalized dictionaries are confronted with great challenge. The all-round convergence between media convergence and dictionary in the aspects of compilation and publication involves sources, texts, modalities, media and users, even compatibility and coexistence between traditional dictionaries and dictionaries based on media convergence as well. So many segments, wide coverage and underlying depth that the traditional dictionaries have never seen before. And this is exactly the primary problem that we discover through field survey for the tendency of digital compilation and publication in recent times. In order to suit the remedy to the case and overcome difficulties, such technologies as digital and Internet communication, media convergence and language, essential points as lexicographical text of multimodalities and entry meta-data, footholds as intelligent reference demand and habit of users are all in urgent need. Besides, how to guarantee the safety of intellectual property of lexicographical text used in digital media among Internet space is also a realistic problem to restrain the development of digitalized dictionaries in the age of media convergence.

5 References

Kilgarriff, A. (2006). If dictionaries are free, who will buy them?. Kernerma Dictionary News, Number 13, June.

Kress, G. & T. can Leeuwen. (2006). Reading Images: The Grammar of Visual Design (2nd edition). London & New York: Routledge.

Leminen, S., Huhtalta, J., Rajahonka, M. et al. (2016). Business model convergence and divergence in publishing industries. In A. Lugmayr & C. Dal Zotto (eds.) Media Convergence Handbook. Vol. 1 (pp. 187-202). Heidelberg, New York, Dordrecht, London: Springer.

Lew, R. (2020). Multimodal lexicography: The representation of meaning in electronic dictionaries. Lexikos, 20, 290-306.

Veglis, A., Dimoulas, C., Kalliris, G. (2016). Towards intelligent cross-media publishing: Media practices and technology convergence perspectives. In A. Lugmayr & C. Dal Zotto (eds.) Media Convergence Handbook Vol. 1 (pp. 131-150). Heidelberg, New York, Dordrecht, London: Springer.

Chen, Wei. (2010). Dictionary publishing industry in the digital age. Publishing Journal, 6, 86-90. Dai, Yuanjun. & Xu Hai. (2014). Electronic lexicography: Status quo and prospects. Lexicographical Studies, 4, 1-9.

Liu, Chengyong. (2007). An analysis of "dictionary corpus and compilation system launched by The Commercial Press". Science and Publication, 12, 19-20.

Lv, Jing. (2020). Digital age: Challenge for paper dictionaries and opportunity for dictionaries based on media convergence. View on Publishing, 13, 42-44.

Zhang, Yihua & Liu, Hui. (2007). Bilingual dictionary generation system based on micro-data structure. Foreign Language and Their Teaching, 8, 61-64.

Zhang, Yihua. (2019). On the innovation of dictionary compilation and publication in the context of media convergence. Chinese Journal of Language Policy and Planning, 6, 79-89.

Zhang, Yihua. (2021). Design conception of multimodal-lexicographical texts from the perspective of convergent media. Lexicographical Studies, 2, 20-32.

Zhou, Zhixin & Bai Lijing (2014). Developmental model for dictionary publishers in the age of digital publication. View on Publishing, 2, 71-73.

Zhuang, Zhixiang., Zhang, Chunming., Zhang, Yihua. (2013). Research and Development of Bilingual Dictionary Compilation System. Shanghai: Shanghai Scientific & Technical Publishers.

INSCRIBING IDENTITY AND SEARCHING FOR THE ORIGIN: COMPILING AND BUILDING A DICTIONARY OF MALAY INSCRIPTIONS

Totok Suhardijanto, Ninny Soesanti
Faculty of Humanities, Universitas Indonesia
suhardiyanto@gmail.com; niniesusanti@gmail.com

Abstract

The civilization of scriptwriting in Malay language has started with the discovery of Kedukan Bukit Inscription originating from the 683 CE in South Sumatra. Since then, many inscriptions written in Malay occurred and found in the Mainland and Islands of Southeast Asian including the Malay peninsula, Luzon, Sumatra, and Java. They are written on stones and copper plates. Until now, the lexicographic study of Malay inscriptions still lags if we did not want to say never been studied. This paper presents our effort to document and conserve the Old Malay texts written in the inscriptions. In this paper, Old Malay refers to a variety of Malay used in tens of inscriptions that were created around 7 and 14 the century. We used the terms of "inscription" in its widest meaning. Many people believe that Old Malay has inadequate source for linguistic studies and lexicographic works. Data for this project are collected from several inscriptions in the Mainland and Islands of Southeast Asia. For lemmatization, we used an etymological and historical linguistic method to determine whether a word should be included or not in the first edition of this dictionary. Definition and gloss were created by consultation with reliable references such as R.J. Wilkinson's Dictionary and Klinkert's Dictionary of Malay to consider whether a word is or not a Malay word. For the lexicographic work, we make use of Lexonomy software that is very practical and user-friendly. The result shows that the building of Old Malay that is considered a "dead" language has challenges of the limited surviving evidence and the needs of users.

Keywords Old Malay, Malay inscriptions, Dictionary of Malay Inscriptions.

1 Introduction

The history of Malay has begun since the Kedukan Bukit inscription dated from 683 CE was discovered by the Dutchman C.J. Batenburg on 29 November 1920 near Palembang, South Sumatra. This inscription was written in Pallava script (De Casparis 1978). Pallava script is a Brahmic script, named after the Pallava dynasty of South India, attested since the 4th century AD (Griffiths 2014). The Kedukan Bukit inscription is the oldest specimen of the Malay language, known as Old Malay. According to Griffiths (2018), Old Malay can be defined as the variant of the Malay language found in documents written in an Indic (i.e., Brāhmī-derived) system of writing. Most of Old Malay inscriptions were written in Pallava script. Only one, that is the Tanjung Tanah manuscript, which was written in Old Sumatra script, a local adopted and modified script derived from Brahmi script.

This paper presents our effort in documenting and building a dictionary of a language in what is called Old Malay. In this paper, we prefer using the language of Malay inscriptions to Old Malay because of several reasons. First, it is used to avoid what Griffith (2018: 275) mentioned: "… a negative definition, aiming to capture the state(s) of the Malay language before it had undergone any influence from Arabic and Persian … all of which are marked by a significant percentage of Arabic loan vocabulary." Second, the number of text sources available in Old Malay is still very limited. It remains difficult to carry on a comprehensive study on Old Malay. Third, Old Malay is dialectally, not uniform.

To date, dozens of Malay inscriptions are found in the mainland and the island of Southeast Asia, including the Malay Peninsula in Malaysia and Thailand, Sumatra, Bangka, Java, and Borneo in Indonesia, and Luzon in the Philippines. However, the use of the "inscription" term in this paper is not fully appropriate since one of the Old Malay documents from Kerinci is in a form of a manuscript. According to Kozok (2004: 39), this manuscript widely known as Tanjung Tanah Manuscript may prove to be the oldest extant Malay language manuscript. Before this manuscript, Old Malay texts were only written in unperishable material such as stones, plates, and animal horns. Unlike other Old Malay manuscripts which were written in Pallava script, the Tanjung Tanah Manuscript was written in what Voorhoeve (1970) calls "Old Javanese" which is one of the local Sumatran Late Pallavo- Nusantara scripts (Kozok 2004: 39), which Casparis (1975: 57) more aptly called "Malayu".

2 Method

Data for this lexicographic work are collected from several inscriptions in the Mainland and Islands of Southeast Asia. According to Utomo & Shuhaimi (2009), 51 inscriptions are recorded and widely recognized as Malay inscriptions. However, for this paper, corpus data are restricted to a collection of three text sources, that is the Kedukan Bukit Inscription (KBI), the Talang Tuwo Inscription (TTI), and the Tanjung Tanah Manuscript (TTM).

KBI was discovered by the Dutchman C.J. Batenburg on 29 November 1920 at Kedukan Bukit, South Sumatra, Indonesia, on the banks of Tatang River (Bloembergen, M. and Eickhoff 2020). The inscription is known as the oldest surviving specimen of the Malay language, in a form known as Old Malay (Guy 2014). It is in a shape of a small stone of 45 cm × 80 cm and written in Pallava script. Based on the analysis of radiocarbon, this inscription is dated 1 May 683 CE.

According to Cœdès (1968), TTI is a 7th-century Srivijaya inscription in a size of 50 cm x 80 cm discovered by Louis Constant Westenenk on 17 November 1920, on the foot of Bukit Seguntang near Palembang. The inscription was discovered in good condition with clearly inscribed scripts. It was made from a stone block and dated from 23 March 684. The inscription was written Pallava script in Old Malay.

Kozok (2014) described that TTM is hence considerably older than the previously oldest known Malay manuscripts. As a Malay manuscript written in a Late Pallavo-Nusantara script, TTM indicates that there was a tradition of Malay writing on perishable material that predates the introduction of Muslim and European paper and Jawi script and suggests that this tradition may extend back as far as the earliest Malay inscriptions in the 7th century (in Pallava script). TTM manuscript also makes obsolete the theory that there was no tradition in the Malay world of writing on palm leaf or similar materials before the arrival of Islam (Jones 1986: 139). Based on the radiocarbon date and the available historical sources, it seems to be highly probable that the manuscript dates to the 14th century.

All linguistic data from two inscriptions and one manuscript were collected for lemmatization. In the next step, we used an etymological and historical linguistic method to determine whether a word should be classified as Malay or non-Malay origins. Definition and gloss were created by consultation with reliable references such as R.J. Wilkinson's Dictionary and Klinkert's Dictionary of Malay. For the lexicographic work, we make use of Lexonomy software (Měchura 2017) that is very practical and user-friendly. With Lexonomy, everyone, individuals, and teams can create a dictionary, design an arbitrary XML structure for the entries, edit entries, and eventually make the dictionary publicly available as a 'microsite' within the Lexonomy website. Moreover, Lexonomy also provides users with the functionality to publish a print version of their works.

3 Result and Discussion

In this session, we discuss several issues regarding the writing process of this Old Malay dictionary. These issues comprise of the circumstances of Old Malay corpora, the software tools for querying corpus and writing dictionary, and the entry system of the dictionary.

3.1 The Sources of Old Malay Texts

Regarding the compiling process of a dictionary of Old Malay inscriptions, there are several matters that should be explained here. As mentioned before, one of the difficulties in compiling a dictionary of Malay inscription is that the source texts of Old Malay are very limited. Not including TTM, there are only 51 inscriptions found in Southeast Asia, mainly in the region of Indonesia. Table 1 shows dozens of inscriptions date around 7th century. Most of the texts are ranging from one sequence of words to dozens of lines of sentences.

The second problem is the condition of stones. Some texts were partially transcribed because their stone materials had eroded. A small number of texts is still unreadable at all due to their damaged materials. Among the texts included in Table 1, Kedukan Bukit, Talang Tuo, Koto Kapur, Telaga Batu, and Karangberahi are among the transcriptions with texts in good condition.

Table 1 List of Old Malay Inscriptions from 7th Century

Name of Inscriptions	Year	Region	Sources
1. Kedukan Bukit	682	Palembang, South Sumatra	Ronkel (1924:19-21), Çœdès (1930:33-31 #l), Ferrand (1932:273), Poerbatjaraka (l952: 33 -34), Suhadi (1983:76)
2. Talang Tuo	684	Talang Kelapa, South Sumatra	Ronkel (192412-19), Çœdès (193038a4 #2), Ferrand (1932:276-277), Poerbatjaraka (1952: 35-38), Suhadi (1983:76-77)
3. Koto Kapur	686	Bangka	Kern (1913), Çœdès (1930:46-50 #4), Ferrand (1932:280-281), Poerbatjaraka (1952: 39l), Suhadi (1983:17)
4. Telaga Batu	7th century	Palembang, South Sumatra	Casparis (1956:1 1-15 #le)
5. Karangberahi	7th century	Merangin, Jambi	Krom (1920:426-431 #XVD, Çœdès (1930:45 #3), Boechari (1979), Suhadi (1983:78)
6. Palas Pasemah	7th century	South Lampung, Lampung	Boechari (1979), Suhadi (1983:78-79)
7. Bungkuk	7th century	East Lampung, Lampung	Utomo & Suhaimi (2009)
8. Boom Baru	7th century	Palembang, South Sumatra	Atmodjo (1992)
9. Siddhayātra D 156	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
10. Siddhayātra D 157	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
11. Siddhayātra D 158	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
12. Siddhayātra D 159	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
13. Siddhayātra D 160	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
14. Siddhayātra D 161	7th century	Palembang, South Sumatra	Stutterheim (1936), de Casparis (1956)
15. Kambang Purun I	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
16. Siddhayātra Kambang Purun II	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)

17. Siddhayātra Kambang Purun III	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
18. Kambang Purun IV	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
19. Kambang Purun V	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
20. Siddhayātra Kambang Purun VI	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
21. Siddhayātra Kambang Purun VII	7th century	Palembang, South Sumatra	Utomo & Suhaimi (2009)
22. Siddhayātra Kambang Unglen I	7th century	Palembang, South Sumatra	Atmodjo (1993)
23. Siddhayātra Kambang Unglen II	7th century	Palembang, South Sumatra	Kartakusuma (1993)
24. Siddhayātra	7th century	Tapin, South Kalimantan	Nastiti (1998)
25. Bukit Siguntang D 164	7th century	Palembang, South Sumatra	Casparis (1956:2-6 # 1a)
26. Bukit Siguntang	7th century	Palembang, South Sumatra	Casparis (I956:2-6 # 1a)
27. Siddhayātra D 126	7th century	Bangka	Krom (1926) Cœdes (1989)

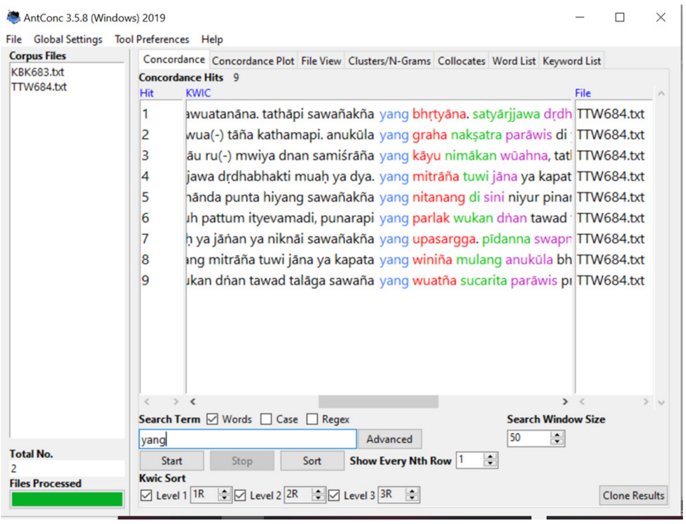
The other problem is that although most of Old Malay inscriptions were written in Pallava, some texts were written in other writing systems. For instance, Sang Hyang Wintang inscription in Temanggung, Central Java, was written in Old Javanese script and TTM (Tanjung Tanah Manuscript) in Old Sumatra script. Furthermore, although many texts were written in Pallava, there are dialectal differences among the scripts.

In addition to the scripting dialect, dialectal differences are also shown in the lexical, grammatical, and discursive elements. Mahdi (2005: 182) stated that several inscriptions (Kota Kapur, Karang Brahi, Palas Pasemah, and the Sabokingking Naga stone) have a non-Old Malay introductory formula. Its language, called ‘language B’ by Damais (1968), bears similarities with Malagasy (Aichele 1954, Damais 1968, Adelaar 1989:36-37). According to Mahdi, three long ones, Karang Brahi, Palas Pasemah, and Kota Kapur are practically identical. The fragmentary Sabokingking B and incomplete Kedukan Bukit inscriptions represent partly overlapping passages of the same text. Old Malay texts abound with Sanskritisms retaining original Sanskrit spelling. Griffith (2018: 279) wrote that in Gandasuli inscription (date 827 CE), besides several lexical items that are a unique or very rare grammatical phenomenon that is of great importance for the history of Malay language, namely the use of vatu = batu as a numeral classifier.

Despite the limited sources of texts in Old Malay, a dictionary of Old Malay inscriptions needs to be made because of these reasons. First, it is very rare to find out work with systematic and comprehensive descriptions of Old Malay (Griffiths 2018, Mahdi 2005, Kozok 2004). Second, up to now, the discoveries of Old Malay inscriptions continue to occur. For example, in recent years, the discovery of substantial numbers of inscribed tin foils from the Batang Hari River at Muara Jambi, and reportedly now also from the Musi at Palembang, has revealed an almost entirely new genre of inscribed materials that are part of the Old Malay corpus (Griffiths 2018). Third, a systematic and comprehensive collection of Old Malay texts will reveal the relationship between Old Malay and other Malay variants, such as Classic Malay or Modern Malay, and will help to explain “the missing link” in the history of Malay.

3.2 Software Tools

This section describes the features and benefits of both types of programs, namely the corpus query system, and the dictionary writing system. As for the first type, we implemented two different corpus query systems. First, AntConc version 3.5.8 (Anthony 2019) was used when we wanted to look up a word used in contexts through concordance functionality. Second, since we are working with an ancient language, we need to know the translation of the text. For this reason, we also make use of software that can deal with parallel corpus. In this case, we chose AntPConc version 1.2.1 (Anthony 2017) whose functionalities to manage multilingual corpora.


AntConc 3.5.8 (Windows) 2019
File Global Settings Tool Preferences Help
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Advanced
Start Stop Sort
Clone Results

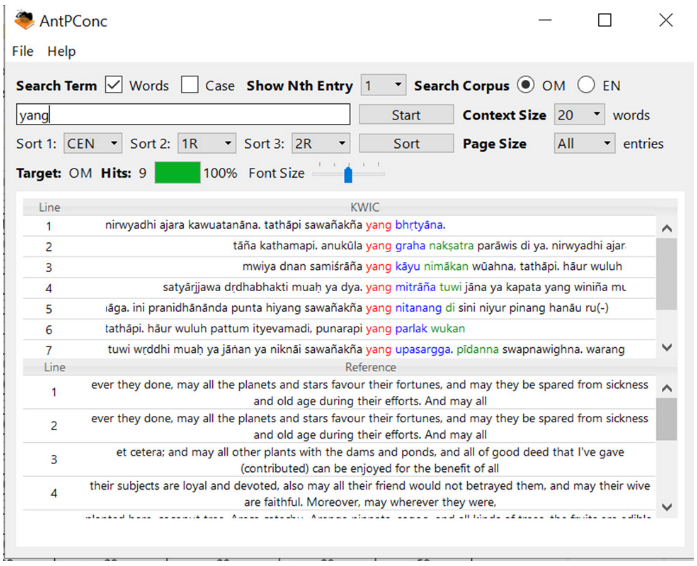
Corpus Files
KBK683.txt
TTW684.txt

Concordance Hits 9

Hit	KWIC	File
1	awuatanāna. tathāpi sawañakña yang bhṛtyāna. satyārjjawa dṛdh	TTW684.txt
2	wua(-) tāña kathamapi. anukūla yang graha nakṣatra parāwis di	TTW684.txt
3	āu ru(-) mwiya dnan samiśrāña yang kāyu nimākan wūahna, tatl	TTW684.txt
4	jawa dṛdhabhakti muaḥ ya dya. yang mitrāña tuwi jāna ya kapat	TTW684.txt
5	nānda punta hiyang sawañakña yang nitanang di sini niyur pinaı	TTW684.txt
6	ıh pattum ityevamadi, punarapi yang parlak wukan dṅan tawad	TTW684.txt
7	ḥ ya jāṅan ya niknāi sawañakña yang upasargga. pīdanna swapn	TTW684.txt
8	ıng mitrāña tuwi jāna ya kapata yang winiña mulang anukūla bh	TTW684.txt
9	ıkan dṅan tawad talāga sawaña yang wuatña sucarita parāwis pı	TTW684.txt

Search Term ☑ Words ☐ Case ☐ Regex
yang
Search Window Size 50
Show Every Nth Row 1
Total No. 2
Files Processed
Kwic Sort ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 1: KWIC concordance for the relative pronouns *yang* in the AntConc



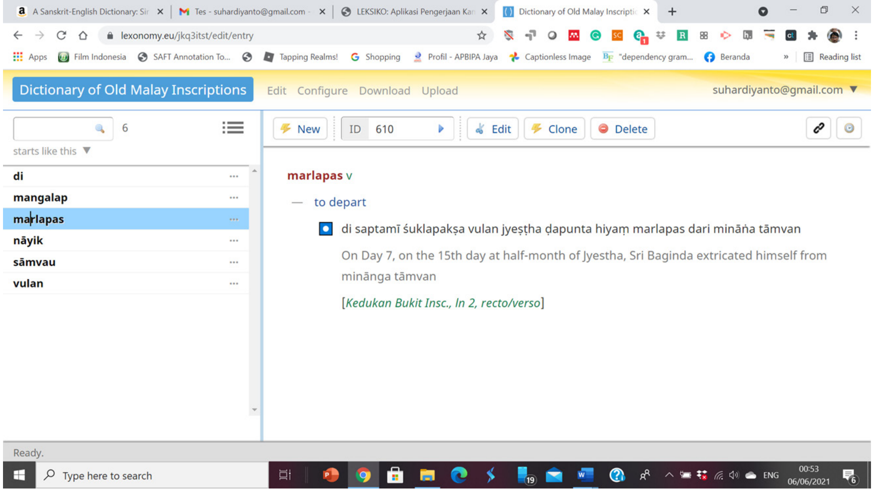
Search Term ☑ Words ☐ Case Show Nth Entry 1 Search Corpus ● OM ○ EN
yang
Context Size 20 words
Sort 1: CEN Sort 2: 1R Sort 3: 2R
Page Size All entries
Target: OM Hits: 9 100%

Line	KWIC
1	nirwyadhi ajara kawuatanāna. tathāpi sawañakña yang bhṛtyāna.
2	tāña kathamapi. anukūla yang graha nakṣatra parāwis di ya. nirwyadhi ajar.
3	mwiya dnan samiśrāña yang kāyu nimākan wūahna, tathāpi. hāur wuluh
4	satyārjjawa dṛdhabhakti muaḥ ya dya. yang mitrāña tuwi jāna ya kapata yang winiña mu
5	ıāga. ini pranidhānānda punta hiyang sawañakña yang nitanang di sini niyur pinang hanāu ru(-)
6	tathāpi. hāur wuluh pattum ityevamadi, punarapi yang parlak wukan
7	tuwi wṛddhi muaḥ ya jāṅan ya niknāi sawañakña yang upasargga. pīdanna swapnawighna. warang

Line	Reference
1	ever they done, may all the planets and stars favour their fortunes, and may they be spared from sickness and old age during their efforts. And may all
2	ever they done, may all the planets and stars favour their fortunes, and may they be spared from sickness and old age during their efforts. And may all
3	et cetera; and may all other plants with the dams and ponds, and all of good deed that I've gave (contributed) can be enjoyed for the benefit of all
4	their subjects are loyal and devoted, also may all their friend would not betrayed them, and may their wive are faithful. Moreover, may wherever they were,

Figure 2: KWIC concordance for the relative pronouns *yang* from Old Malay texts and English translation in the AntPConc

The second type of works is writing and compiling a dictionary. For this purpose, we need software that can help us to manage all issues regarding lexicographic works. In this project, we chose Lexonomy (Měchura 2017) which is introduced to the public at Electronic Lexicography Conference (ELEX) 2017 in Leiden, the Netherlands. This software has a mission to be an easy-to-use tool for small to medium-sized dictionary projects. In Lexonomy, individuals and teams can create a dictionary for any purpose. Once a dictionary has been created, the user who first created it may add additional users, and then they can all start adding and editing entries. Měchura (2017) stated that Lexonomy has features related to three stages of the writing dictionary process, namely dictionary planning: specifying the structure of entries, etc, dictionary editing: adding and deleting items, and online dictionary publishing: generating an electronic version of the dictionary.



di
mangalap
marlapas
nāyik
sāmvau
vulan

marlapas v

— to depart

di saptamī śuklapakṣa vulan jyeṣṭha ḍapunta hiyaṃ marlapas dari mināṅa tāmvan

On Day 7, on the 15th day at half-month of Jyestha, Sri Baginda extricated himself from minānga tāmvan

[*Kedukan Bukit Insc., ln 2, recto/verso*]

Figure 3: Dictionary Editor Functionality in the Lexonomy

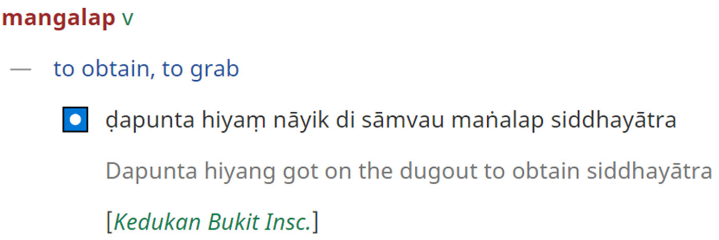
3.3 Entry Structure of the Dictionary

It is very common for Malay or Indonesian dictionaries to be made in the form of morpheme-based dictionaries. For this reason, users of Malay/Indonesian dictionaries require knowledge of morpheme identification to find out the appropriate meaning of a word they look for. However, about this dictionary, we decided to compose it in a word-based dictionary because of scarce data in Old Malay texts. It is slightly reckless to determine the morpheme of a word based on insufficient data.

Regarding types of entry, entries in the Dictionary of Old Malay Inscription are managed in a word-class-based organization. According to Atkins and Rundell (2008), at least there two main reasons for choosing a word-class-based organization. First, it is the more usual way of handling dictionary entries, so most users will be familiar with it. Second, as an arbitrary access system (like alphabetical order), it can be applied objectively and systematically.

About the entry structure, senses are ordered based on the frequencies in the corpus. By this system, the most frequent sense will be placed at the first order. However, due to data shortage, it is very rare to have a lemma with more than one sense. At the moment, for this paper, the dictionary only provides an English translation for each sense. Soon, an Indonesian translation version will be added.

Examples are included for each sense within the entry of one headword. For example, it is looked up from the corpus and be accompanied by an English translation. In addition to English translation, information of the location where the inscription contained the text found is also provided. Moreover, for further studies, especially in inscription and manuscript studies, information about lines, inscription sides, and manuscript pages are also provided like in philology and epigraphy works. Figure 5 shows the layout of entry marlapas in the dictionary.



mangalap v

— to obtain, to grab

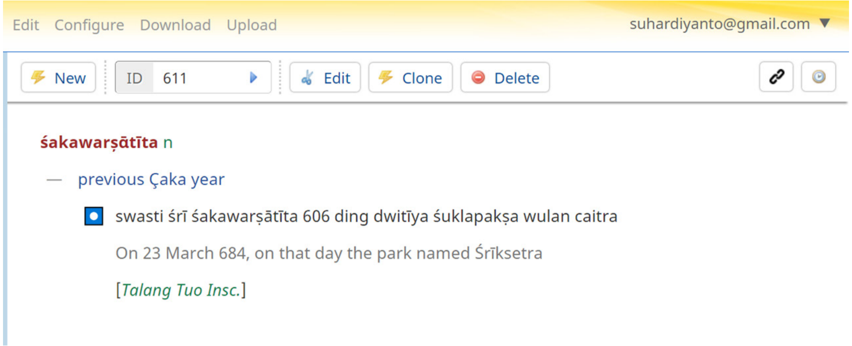
ḍapunta hiyaṃ nāyik di sāmvau maṅalap siddhayātra

Dapunta hiyang got on the dugout to obtain siddhayātra

[*Kedukan Bukit Insc.*]

Figure 5: The entry for *mangalap* in the dictionary

In many cases, Old Malay inscriptions contained not only Malay words but also Sanskrit words. For older inscriptions, Sanskrit words even occur more frequently than those of Malay. For this reason, Old Malay inscriptions also contained many MWEs in Sanskrit. According to Atkins and Rundell (2008), usually MWEs are treated in three ways. First, they are included within the entry of one of their component words. Second, they are treated as secondary headwords or they may be put in a separate section of the entry. Third, they are treated as a separate entry distinct from any related entry. In this dictionary, MWEs are treated as a separate entry like in the example in Figure 6.



śakawarṣātīta n

— previous Çaka year

● swasti śrī śakawarṣātīta 606 ding dwitīya śuklapakṣa wulan caitra

On 23 March 684, on that day the park named Śrīksetra

[*Talang Tuo Insc.*]

Figure 6: The entry for MWE śaka***warṣātīta*** **in the dictionary**

4 Conclusion

This paper presents an attempt in documenting and creating a dictionary of a language with scare data which is called Old Malay. In the discussion section, we explained how the corpus data of Old Malay and why this dictionary need to be created. In this paper, we only limited our corpus to three inscriptions. Unlike other Malay/Indonesian dictionaries ever created, this dictionary was made with a word-based organization. The dictionary tried to keep the form as it is following the result of authorized scholars' reading and transcription. Types of entry are organized on a word-class basis. Senses are set in frequency order where the most frequent sense is put in the first order. An English translation is provided for each definition and example. In addition to the example, the dictionary also provides us with information about the location where the inscription is found and where the example text is located in the source, either inscription or manuscript.

5 References

Anthony, L. (2017). AntPConc (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Ashdowne, R. (2016). Dictionaries of Dead Languages. In Phillip Durkin (Ed.). Oxford Handbook of Lexicography. Oxford: Oxford University Press.

Atkins, B.T.S. & Rundell, M. (2008). The Oxford Guide: Introduction to Practical Lexicography. Oxford: Oxford University.

Bloembergen, M. and Eickhoff, M. (2020). The Politics of Heritage in Indonesia: A Cultural History. Cambridge University Press. ISBN 978-1-108-49902-6.

Bloembergen, M. and Eickhoff, M. (2020). *The Politics of Heritage in Indonesia: A Cultural History*. Cambridge University Press. ISBN 978-1-108-49902-6.

Casparis, J.G. de, 1975, *Indonesian palaeography : a history of writing in Indonesia from the beginnings to c. A.D. 1500, Handbuch der Orientalistik: 3. Abt., Indonésien, Malaysia und die Philippinen; 4. Bd. Linguistik, 1. Lieferung*, Leiden : E.J. Brill.

Cœdès, George (1968). Walter F. Vella (ed.). The Indianized States of Southeast Asia. trans. Susan Brown Cowing. University of Hawaii Press. ISBN 978-0-8248-0368-1.

De Casparis, J. G. (1978). Indonesian Chronology. *BRILL Academic*. pp. 15–24. ISBN 90-04-05752-8. Fuertes-Olivera, P.A. (2017). *The Routledge Handbook of Lexicography*. Amsterdam: Routledge Publishing. Griffiths, A. (2014). Early Indic Inscriptions of Southeast Asia. in J. Guy (Ed). *Lost Kingdoms: Hindu- Buddhist Sculpture of Early Southeast Asia*. New York: Metropolitan Museum of Art. Hlm, 53-57. Griffiths, A. (2018). The Corpus of Inscriptions in the Old Malay Language. In D. Perret (Ed). Writing for Eternity: A Survey of Epigraphy in Southeast Asia. Paris: École française d'Extrême-Orient, pp. 275— 286.

Guy, J. (2014). Lost Kingdoms: Hindu-Buddhist Sculpture of Early Southeast Asia. Metropolitan Museum of Art. p. 21. ISBN 9781588395245.

Kozok, U. (2004). A 14 th Century Malay Manuscript from Kerinci. Archipel, 67(1), 37-55.

Měchura, M. B. (2017) 'Introducing Lexonomy: an open-source dictionary writing and publishing system' in Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.

Monier-Williams, M. (2005). A Sanskrit-English Dictionary. Bharatiya G.N. (Educa Books) .

Perret, D. (2018). Writing for Eternity: A Survey of Epigraphy in Southeast Asia. Paris: École française d'Extrême-Orient.

Suhardijanto, T. and Dinakaramani, A. (2017). Building a Collaborative Workspace for Lexicographic Works in Indonesia. in Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.

Susanti, Ninny. (2019). Script and Identity of Indonesia. *MALINDO-Journal of Malaysian and Indonesian Studies*, Volume 1(1), October 2019, 1-7.

Utomo, B.B. and Shuhaimi, N. H. (2009). Inskripsi Bahasa Malayu Kuno di Asia Tenggara. Kuala Lumpur: Universiti Kebangsaan Malaysia.

Voorhoeve, P. (1970). Kerintji Documents, *Bijdragen tot de Taal-, Land- en Volkenkunde*, 126 (4) 369 399 .

KAMUS BAHASA JAWA BANYUMASAN-INDONESIA: A STRATEGY OF SHOWING LEXICOGRAPHIC EVIDENCE TO MAINTAIN A LOCAL WISDOM

Tri Wahyuni
Regional Agency for Language in Central Java Province, Indonesia
tri.wahyuni@kemdikbud.go.id

Abstract

One of the efforts to make a language and cultural codification is the dictionary making. The main reasons why people open the dictionaries is that they want to get many information about the meaning. *Kamus Bahasa Jawa Banyumasan—Indonesia* is the one of bilingual dictionary as a lexicographic evidence based on dialect in Central Java that show the local wisdom by its entries. The uniqueness of culture in Banyumas which has a local wisdom was a part of entries in it. There were several strategies of the lexicographer to arrange the entries of this dictionary i.e. (1) selective in a collecting the data, (2) put the proper definition, (3) knowledge about the local culture, (4) identifying the senses, and (5) put a proper microstructure of the dictionary. The lexicographer's strategy of dictionary making decide the quality of dictionary to show all the local wisdom in the particular community, especially in Banyumas dialect. Therefore, a complete explanation of the meanings of lexical items in this dictionary is the main thing to make in order to show a natural local wisdom.

Keywords codification, local wisdom, lexicographic evidence, lexicographer's strategy

1. Background

The dictionary can be said to be one of the real forms of language and cultural codification. Through a dictionary, a person will be able to describe certain language and cultural conditions. Dictionaries are generally used when people want to know the meaning of a word in certain languages and cultures. The purpose of making the dictionary, of course, also varies, depending on the insights and motivations of the lexicographer. This article discusses the existence of the *Kamus Bahasa Jawa Banyumasan-Indonesia* (Banyumasan-Indonesian Javanese Dictionary)—and then abbreviate to be KBJBI—, which is a bilingual dictionary based on the Banyumas dialect in Central Java. The dictionary can be said to be proof or evidence of lexicographical work carried out for one important purpose, in this case to maintenance *Banyumasan* local wisdom.

This article intends to illustrate that the dictionary can be used as evidence of defending local wisdom. Component analysis can be used to reveal the local wisdom contained in a language. The component analysis in question can sort out a linguistic unit into the elements that reconstruct it. Apart from component analysis, there is also semantic analysis which is used to obtain the meanings of words in a particular language. According to Putra (2009) there are two types of wisdom related to culture, namely traditional wisdom and local wisdom. Even though they are both wisdoms, there are differences in definitions between traditional wisdom and local wisdom.

Traditional wisdom is 'a set of knowledge and social practices to solve problems and/or difficulties faced in a good and right way. Local wisdom is 'a set of knowledge and practices in a community, both from previous generations and from experiences related to the environment and other communities to solve problems whether they have legal force or not. So, the KBJBI which lists the local wisdom of the

Banyumasan people must be seen in the form of the context by lexicographer who has knowledge of local culture in defining the entries listed in it.

In the past, the dictionary was defined as a form of habit arranged systematically as expressed by Zgusta (1971: 17 in Sterkenburg, 2003: 4).

> "A dictionary is a systematically arranged list of socialized linguistics forms compiled from the speech habits of a given speech community and commented on by the author in such a way that the qualified reader understands the meaning…of each separate form, and is informed of the relevant facts concerning the function of that form in its community."

The Swedish lexicographer, Bo Svensén (1993: 3-4 in Sterkenburg, 2003: 4) provides some explicit definition of dictionary. He argued that a dictionary is a book that in the first place contains information on the meaning of words and their usage in specific communicative situations. The dictionary is examined here as a work in its own right although it is just one component of descriptive work. It was also a work of its time and context[1]. Henderson also said that the collection of text is particularly significant because texts can provide as wide range of words and expressions used in a range of context.

Javanese language is generally known as a regional language which has experienced significant development. Javanese is a regional language that is included in the Austronesian language family, West Polynesian Malay which is known to have a fairly large number of speakers. Based on Hidayat and Rahmani's research (in Paryono, 2011) the number of Javanese speakers is ranked 11th among 6,703 languages in the world, with a percentage of 80 to 100 million Javanese speakers. The wide spread of the use of the Javanese language makes Javanese language develop along with the natural and speaking community conditions. According to Soedjito et al. (in Paryono, 2011) Javanese has several geographic dialects such as the Javanese Banyumas dialect, Tegal, Surakarta/Yogyakarta, Surabaya, Samin, and Osing. The Javanese speaking community has a kind of convention to make the Javanese language spoken in the Yogyakarta and Surakarta regions as standard Javanese. The reason is, perhaps because the Javanese language in this area was inherited by the Javanese monarchy (Mataram kingdom) which came into power after the collapse of Majapahit kingdom.

According to Chaer (2007: 59), what is meant by language society is a group of people (in a relatively large number), who feel they are of the same nationality, same ancestry as their area of residence, or have the same social interests. The Javanese language community is a group of people who feel they use the same language, namely Javanese. The focus of the understanding of the language community is on the same feeling in terms of language use, so the concept of language society can have broad or narrow meanings. For example, the Javanese language community has a broader meaning than the Banyumasan Javanese language community, for example.

The Dutch linguist, E.M. Uhlenbeck (1964) classified Javanese according to three geographic dialect groups, namely the western group consisting of the Banten dialect, the Cirebon dialect, the Tegal dialect, the Banyumas dialect, and the Bumiayu dialect (transitionalTegal and Banyumas). The Tegal, Banyumas, and Bumiayu dialects are often referred to as the Banyumasan Javanese language. There is another middle group consisting of the Pekalongan dialect, the Kedu dialect, the Bagelen dialect, the Semarang dialect, the North East Coast dialect (Jepara, Rembang, Demak, Kudus, Pati), the Blora dialect, the Surakarta dialect, the Yogyakarta dialect, and the Madiun dialect. The second group is also known as Middle Javanese or mataraman, which in its development, the Surakarta and Yogyakarta dialects became the standard reference for the official use of Javanese (standard Javanese). Then there is the eastern group consisting of the East Javanese pantura dialect (Tuban, Bojonegoro), the Surabaya dialect, the Malang dialect, the Jombang dialect, the Tengger dialect, and the Banyuwangi dialect or what is often called the language of Javanese wetanan.

[1] Henderson, John. (No Year). Assesing Carl Strehlow's Dictionary as Linguistic Description: Present Value and Future Potential. URL: https://www.jstor.org/stable/j.ctv5cgb2c.13

The dialect of Yogyakarta-Surakarta, which is included in the middle dialect, has become the standard language because of the monarchy system, namely the Mataram kingdom, which is located in these two regions. Meanwhile, other dialects in Java are small dialects that are only spoken in the regional area of the language, one of which is the Banyumas dialect which is then called the Javanese Banyumasan language. People outside the Banyumas region often call the Javanese language Banyumas accent with the Banyumasan Javanese language or often called the Javanese *Ngapak* language. The Banyumasan Javanese language is a group of Javanese languages spoken in the western part of Central Java. Some of the vocabulary and dialects are also used in North Banten, Cirebon, and Indramayu areas. The accent is slightly different from the standard Javanese accent. This is because the Banyumasan Javanese language is still closely related to ancient Javanese or Kawi language.

Phonologically, the standard Javanese language does have a difference from the Banyumasan Javanese. The main difference is that the suffix ‘a’ in the Banyumasan Javanese language is still pronounced /a/ instead of /o/. The pronunciation of the sound /a/ is closer to the pronunciation of ancient Javanese. So, if in Surakarta and Yogyakarta people eat /sego/ [səgɔ] ‘rice’, in Banyumasan people eat /sega/ [səga^{ʔ}]. In addition, words that end in dead letters are read in full, for example the word /enak/ ‘delicious’ is read in standard Javanese /enak/ [ena^{ʔ}], While in Banyumasan it is read /enak/ with a clear sound of the letter /k/. That is why the Banyumasan language is referred to by people outside Banyumas as *Ngapak* or *Ngapak-ngapak*. This phenomenon shows that local wisdom is still maintained from the Banyumasan dialect.

Historically, the Banyumasan Javanese language has undergone several stages of development, including in the 9th-13th century Banyumasan Javanese became part of the ancient Javanese language. Then, in the 13-16th century it developed into medieval Javanese. In the 16th - 20th centuries developed into the new Javanese language, and in the 20th century until now, as one of the dialects of modern Javanese. These stages of development may have occurred due to the impact of the kingdoms on the island of Java which also gave rise to the development of feudal cultures. The most important thing in the development process is the implication of the emergence of *undha usuk* or the level of standard Javanese based on one’s social status. This can be said to be a political step to legitimize Mataram’s power through language (Moedjanto, 1987: 25). However, the influence of this feudal culture did not really affect the people in the Banyumas region. This may be the cause of the differences between the Banyumasan Javanese language and the Yogyakarta-Surakarta Javanese language which later became the standard Javanese language.

In the course of history, Mataram’s power during the Dutch colonial era also covered the Banyumas area. This eventually led to the term *bandhekan* in the Banyumasan region to represent the standard Javanese style, or so-called *wetanan* language. Ahmad Tohari[2] stated that the Banyumasan Javanese language can be said to be a humanist and populist language because it does not recognize levels like those in standard Javanese. The Javanese language of Banyumasan which is marginalized is also ancient Javanese because it uses the vowel /a/, but

this opinion needs to be proven by research. In addition, M. Koderi, one of the Banyumasan language and culture experts stated that the word *bandhek* is morphologically derived from the word *gandhek* which means ‘messengers’ (people who are ordered/ordered), namely people who were ordered by the king of Mataram who were sent to the Banyumas region to spread the commandment. the king of Mataram. The messengers of course used the standard Javanese style which was different from the Banyumasan Javanese language which was still influenced by the medieval Javanese style of the Majapahit Kingdom. Power became a tool to force the Banyumas people at that time to also use the language style brought by these *gandhek*. So, the *krama* language that is sometimes spoken by the people in the Banyumas region is the impact of the arrival of the *gandhek* which introduced the level of speech to the Banyumas people who at first did not recognize the level of speech in the language. It is also listed in the KBJBI.

[2] Private deep interview in 2nd and 3rd September 2013 (Wahyuni, 2014)

agem (krm) *v* pakai; **ngagem** *v* memakai; **ageman** *n* pakaian

ageng (krm) *a* besar. *Anaké sampéyan empun --?* Anakmu sudah besar?

As in Javanese in general, in the Javanese Banyumasan language there are values of very noble cultural local wisdom as the identity of the people in the Banyumas region. The uniqueness possessed by the Banyumas people through their language is the culture of speaking frankly, as it is and there is nothing to hide. This culture is known as *cablaka, thok melong*, and *blaka suta*. Etymologically, *blaka suta* comes from the word *blak* or *blag* which means "frankly", to speak as is without further ado. According to Mardiwarsito (1979, in Paryono, 2011) the word *blaka* comes from ancient Javanese, *balaka* and the Sanskrit *walaka* which means straightforward, honest, straight, without being covered up. Meanwhile, the word *suta* means 'child'. When the two words are combined, it will form the meaning of speaking frankly like a child who is still pure, innocent, and what it is. A culture like this is certainly considered rude and disrespectful in the area of standard Javanese because it is considered not knowing manners and is considered disrespectful to others.

Like the standard Javanese language, the Banyumasan language also has a peculiarity of the sound system. The peculiarity of the Banyumasan Javanese vowel phoneme is in the pronunciation [a] of the word /rika/ 'you', it is still pronounced as is [rika], the pronunciation of [i] in the word /isin/ 'shame' is pronounced [isin], the pronunciation of [u] is the word /cimplung/ remains pronounced [cimpluŋ]. The peculiarities of the consonant phonemes that exist in the Javanese language Banyumasan include the phoneme [b] in the word /ababe/ 'air that comes out of the mouth' which is pronounced [abhabe], [d] /babad/ 'tebas' pronounced [bhabhad], [g] /endhog/ 'egg' is pronounced [əndhɔg], [k] /pitik/ 'ayam' is pronounced [pItIk], and the glottal sound [ˀ] /rika'/ 'you' is pronounced [rikaˀ]. In addition, the syllables in Javanese Banyumasan tend to be longer when compared to standard Javanese. This can be seen from the vocabulary /temenan/ [təmənan] 'really', then the word /gemiyen/ [gəmiyɛn] 'first', also the word /gedebogan/ [gədəbɔgan] 'banana tree trunk'.

Along with the rapid development rate of communication and information technology, the existence of the Javanese language Banyumasan can be said to have begun to be shifted by other languages around it. In reality, native speakers of Banyumasan Javanese tend to switch codes when engaging in general communication. In the context of regional communication, Banyumasan Javanese speakers are more dominant in using standard Javanese or *wetanan* languages. One of the basic reasons is that the Banyumasan Javanese language is not a local content material for teaching Javanese in schools. In addition, the formal variety of languages at the government level also uses Javanese manners as a form of respect for speech partners, especially if they are an official or a person who has a higher social stratum than the speaker.

Another fundamental thing that makes Banyumasan Javanese speakers reluctant to use their dialect is a sense of inferiority because Banyumasan Javanese is identified with a variety of coarse languages. Actually, there have been many efforts to maintain the preservation and existence of the Banyumasan Javanese language with the holding of the *Penginyongan* Language Congress in 2016 in Purwokerto, an effort to documenting Banyumasan speech in written and printed form, in the form of books and magazines. However, these efforts need to be carried out massively and continuously. One of the concrete efforts made by a native speaker of Javanese Banyumasan who is also a well-known writer, Ahmad Tohari, was making the *Kamus Dialek Banyumas* (which was later reconstructed by the Balai Bahasa of Central Java Province into the KBJBI) and translating his literary works, such as *Bekisar Merah* translates to *Jegingger* and *Ronggeng Dukuh Paruk* in a local language.

In reality, the efforts to revitalize language through literary works have not been fully maximized. The quantity of literary works using the Banyumasan Javanese language is still very minimal compared to other literary works or publications in standard Javanese. Ahmad Tohari also stated that there are literary works with Banyumas background such as *Babad Kamandaka* which are written in standard Javanese.

Based on these reasons, the author is interested in uncovering the existence of the KBJBI (Banyumasan-Indonesian Javanese Dictionary) as an evidence of lexicographic work and a manifestation of local wisdom in the Banyumas region.

2. Problem

The formulation of the problem in this paper is how the selection of the entries in the *Kamus Bahasa Jawa Banyumasan-Indonesia*? and whether the entries in the *Kamus Bahasa Jaw Banyumasan-Indonesia* are a manifestation of Banyumasan local wisdom?

3 Goals and Benefits

This paper aims to describe the selection of the KBJBI entry and describe local wisdom in the inclusion of entries in the KBJBI. The author wishes that this article will have benefits, both theoretically and practically. Theoretically, this paper is expected to be useful for language development, especially the Javanese Banyumasan dialect. In practical terms, the author wishes that it can be useful for codification efforts and more recent lexicography work, especially in efforts to make a dialect dictionary.

4 Finding and Discussion

This article deals to lexicographic project as an evidence to maintenance a local wisdom, and the object is *Kamus Bahasa Jawa Banyumasan-Indonesia* (KBJBI). Therefore, the writer will show how it works at all entries in it. As a work of documentation of local wisdom, KBJBI is constructed systematically to reach the goal. The lexicographer has to apply the strategy to make the dictionary can be the effort of culture and local wisdom codification. There were several strategies of the lexicographer to arrange the entries of dictionary i.e. (1) selective in a collecting the data, (2) put the proper definition, (3) knowledge about the local culture, (4) identifying the senses, and (5) put a proper microstructure of the dictionary.

4.1 Selective in a Collecting the Data

In general, dictionaries can be divided into two categories, namely prescriptive dictionaries and descriptive dictionaries (Crawforth, 2014). Prescriptive dictionary is a dictionary that contains standard entries and is usually based on certain rules in a language. Meanwhile, a descriptive dictionary is a dictionary that contains entries or entries in the use of vocabulary and terms naturally existing in a particular language. KBJBI is the one example of descriptive dictionary.

A dictionary that has a function as language documentation is not fully realized in the history of language in Indonesia (Kridalaksana, 2003: 165). Basically, the existence of a dictionary is not only required to have a common function, as a reference book, but must be able to completely describe the meaning of the existing entries, including the history of their development. History shows that in the past, the Banyumas region was also under the rule of Mataram which had a high literacy rate. However, there is no historical evidence to suggest the development of dictionaries in this area. The dictionary making effort also seems straightforward without good data updating.

KBJBI consist of entries based on the fact of community's communication context. The data taken from the native speaker who make the dictionary at first. Ahmad Tohari, a well- known writer in Central Java, even in the world, compiled the word in *Banyumasan* dialect to be a simple dictionary, *Kamus Dialek Banyumas,* in 2007. Then, Balai Bahasa Provinsi Jawa Tengah rearrange to make the dictionary based on the lexicographic method. However, there are some things that are not exist in the dictionary, especially in the microstructure section. For example, pronunciations have not been included in the dictionary due to limited human resources or lexicographer knowledge. The lexicographers of the KBJBI are not a native speaker of the Javanese Banyumasan dialect. However, this does not mean that it prevents the continuation of reconstructing the *Kamus Dialek Banyumas* so that it can be more lexicographically "readable".

The main data source comes from the *Kamus Dialek Banyumas* (2007) made by Ahmad Tohari and friends. In addition, the authors also added data sources from various publication sources, including magazines, literary works, and other sources taken from the internet. The data is then sorted in detail, which ones can be entered as head words and which ones should be subentries. Efforts to collect data must be done carefully. Lexicographers are required to have sufficient intuition and knowledge in the fields of lexicology and lexicography. Understanding of culture and matters related to linguistics needs to be continuously honed so that the lexicographers' abilities can become more qualified.

Consistency is very important in the preparation of making the dictionary. Lexicographers should be aware that the main problem in selection is consistency (Arnaud, 2019). Lexicographer knowledge of linguistics is important in selecting dictionary data. They must be able to determine which basic form is used as a head word, and which form is affixed, inflected, reduplicated, or which idiom is being the subentry. Regarding local wisdom, data selection in the KBJBI is sufficient. Data selection based on word classes such as nouns, verbs, adjectives, etc. has been tried well. Several entries showing the uniqueness of Banyumas culture have been listed as main entries. For example, noun categories such as *kudhi, lengger, ronggeng, cowong, cimplung*, and *abid* already illustrate the uniqueness of Banyumas that other regions do not have.

4.2 Put the Proper Definition

A good dictionary certainly includes a good definition. Definition is a lexicographer skill to process words that collect meanings. Understanding of the definiendum and definiens must be good so that the definitions included can be comprehensive. The KBJBI can be said to be a lexicographer effort to describe the egalitarian, open minded, and open culture of Banyumasan Javanese speakers. However, standard Javanese language interference is still visible in the dictionary. This has been explained previously because of the influence of feudalism and expansion carried out by the Mataram kingdom. In view of history, the standard Javanese language is one of the political weapons to legitimize power over the people in the *Penginyongan* area. The adagium of *Adoh Ratu Cedhak Watu* is a manifestation of this legalism.

Along with the times, the dictionary media has also changed, from media in the form of books to electronic or digital media in the form of applications. This has an impact on defining the dictionary in general. As noted in the *Macmillan English Dictionary*, the dictionary definition was originally 'a book that gives a list of words in alphabetical order and explain what they mean' underwent a change in definition, that is 'a reference resource which provides information about words ang their meanings, uses, and pronunciations'[3]. This shows that the dictionary can be said to be a manifestation of human civilization and culture. The dictionary is a project that will never be finished, as Rundell (2008: 2) said that all dictionaries are incomplete, and come under the heading 'work in progress'.

In selecting and writing the proper definitions, the lexicography must be fully aware that they were not actually recording the meaning of a particular words, but they must try to suggest within the available space, as many of the aspects of things defined as will recall it to the readers or allow them to form a good idea in their mind (Guralnik, 1958). The definition in KBJBI is an equivalent, but the lexicographer adds the complement explaining in definition in order to make the reader understand about the meaning of entries better. The one of lexicographer strategy in making a proper definition by adding the sentence example and supplementary explanation.

[3] Budiwiyanto, Adi. http://badanbahasa.kemdikbud.go.id/lamanbahasa/artikel/2796/kamus- dalam-perspektif-budaya-material

jabrig *n* sebutan untuk rambut tak terurus. *Wis wujudé ala tambah-tambah rambuté* --. Sosoknya sudah buruk masih ditambah dengan rambutnya yang tak terurus. (KBJBI, 2015:177)

According to Nida (1975: 200) generally there are two different types of factors give rise to discrepancies in meaning: (1) the fundamental diversities in culture and linguistic backgrounds of the source and receptors, (2) their immediate differences of attitude and disposition, with respect to each other, to the content of the message, to the form of the message, or to the circumstances in which the communication takes place. Banyumasan culture have a different with standard Javanese culture, so the lexicographer must have a good knowledge in making a proper definition in the entry based on the setting background.

sintrén *n* kesenian rakyat Cirebon dan sekitarnya dengan penari perempuan sebagai tokoh utama. (KBJBI, 2015: 408)

cowong I *n* jaelangkung; boneka dari tempurung kelapa dan Jerami mirip ondel-onel yg digunakan sbg media ritual

cowongan *n* ritual meminta hujan dengan memainkan cowong, diiringi music calung dan mantra khas Banyumasan (KBJBI, 2015: 94)

The ability of lexicographer making a proper definition in the entri can help the reader understand about the local culture and get much sufficient information.

4.3 Knowledge about the Local Culture

The lexicographers of the KBJBI are not native speakers of the Banyumasan Javanese language. However, in the dictionary making process, the team of lexicographer involved both the native speakers and the early makers of the Banyumasan dictionary. This can also be used as a kind of motivation to understand the culture that exists in the Banyumasan area. Understanding local culture is very important to do so that the definitions included in the dictionary can truly describe the factual situation of existing cultures and customs. If the lexicographers do not have any knowledge of the local culture, it will be very difficult to define the word or lexicons that will be included in the lexicography work in the form of a dictionary. The basic principle that the dictionary must be able to answer itself is very necessary to do so that ambiguousness can be minimized. The lexicographer must have a strategy to increase the horizon of local culture, so the definition in the dictionary can be qualified well.

The uniqueness possessed by the Javanese Banyumasan language really shows the original character of the Banyumas people who emphasize the conditions as they are, without further ado, and without anything being covered up. As a result of this principle, society seems to be insulted in terms of the language it uses because it is considered rude and unethical. Many Banyumas people feel ashamed and inferior when speaking Banyumasan Javanese in a multicultural realm because they feel they have a lower rank than others. There needs to be a deeper understanding of language in addressing this matter. In general, many lexicons in the Javanese Banyumasan language differ from the lexicons in the standard Javanese (Surakarta/Yogyakarta) both morphologically and phonetically. For example, the word /inyong/ in Banyumasan Javanese corresponds to /aku/ in standard Javanese 'I'. Meanwhile, in ancient Javanese there is lexicon /ingwang/ and in Middle Javanese there is lexicon /ingong/. This fact can certainly be used as a basis for thinking that the Banyumasan Javanese language is closer to ancient Javanese apart from the evidence of the pronunciation of vowel phonemes [a], [i], and [u] which tend to be constant (Wahyuni, 2014).

Dictionary can be called as a historical work beside as a descriptive one. It is a kind of record of the language at an earlier point in time. The local community's language— *Banyumasan* dialect— change over period of enormous social change. Apart from being a reference for finding the meaning of a word in a certain language, the dictionary also has a function as a container for gathering certain cultural concepts that describe the life order of the human speakers of that language. For example, in the KBJBI, the inclusion of entries is as follows.

abid *n* seni atau akrobat obor. *Lagi sunat Inyong detanggapna* --. Ketika saya khitan, dirayakan dng akrobat obor. (KBJBI, 2015: 2)

cimplung *n* singkong yang direbus dl nira (KBJBI, 2015: 88)

ledhek II *n* penari perempuan, ronggeng, lengger (KBJBI, 2015: 250)

lengger *n* **ledhek** (KBJBI, 2015: 254)

ronggeng *n* **lengger** (KBJBI, 2015: 377)

The Banyumasan cultural concept depicted in these entries is evidence that the dictionary can be used as a medium for documentation and cultural education as well as an introduction to local wisdom in Banyumas. There is a peculiarity that other regions do not have in the sense of meaning of these entries.

4.4 Identifying the Senses

Hartmann and James (2002: 125) identified sense as 'one of several meanings that can be established for a word or phrase and covered by a definition in reference work. The meaning in a dictionary is quite significant to be included. In a bilingual dictionary, equivalents are the main thing in definition. However, to meet the standard dialect dictionary which is also a manifestation of local cultural wisdom, a sense or "feeling" of meaning must also be in the definition. The specific meaning of certain entries certainly requires a more detailed explanation. Loading sense is the most crucial thing to do. Therefore, lexicographers are required to have adequate intuition and knowledge of the local wisdom of Banyumasan culture.

Identifying of word senses in lexicographic project demands lexicographers have at their disposal (1) their own intuition and knowledge, (2) existing dictionaries and other reference works, and (3) real word occurrences, drawn from traditional quotation files (Moerdijk, 2003: 273). KBJBI still use the conventional way in identifying the senses of entries. Whereas, if the lexicographers use the modern corpora in internet, for instance, it will increase the quality of the dictionary. For making of good sense, lexicographers may have their own strategy. Someone may have the same meanings in their mind, but the way to explain may will vary from another because many factors, like cognitive and stylistic style background. In KBJBI still have a confusing definition or obscurity, especially in the synonym entries.

ledhek II *n* penari perempuan, ronggeng, lengger (KBJBI, 2015: 250)

lengger *n* **ledhek** (KBJBI, 2015: 254)

ronggeng *n* **lengger** (KBJBI, 2015: 377)

When the lexicographers want to give a meaning of cross reference, they must know the sense of the entry by making a taxonomy and semantic field first. Meaning component will help to make a proper sense for the case like that. The dictionary can be defined as one of the media or tools used or used to know and understand the meaning of a word. However, the dictionary has shortcomings, namely that it only sees differences in the meaning of words based on context, but does not differentiate between the fields of meaning between the words contained in them. In addition, other deficiencies of the dictionary include, among other things, a list of words that can be said to be limited and the inclusion of a synonymy form of a word that is not accompanied by a meaning component as a distinction between one word and another (Nida, 1975: 154, 155, 172. According to Nida, meaning component analysis is needed in defining entries in the dictionary. There are three types, namely components of meaning in general, diagnostics, and supplements. The general component referred to by Nida can be described as a component contained by the meaning of a number of words that are covered in a certain field of meaning. The general component can be said to be a semantic component contained in the same meaning of a number of words in one domain. Apart from the general component, there is a diagnostic component which is also often called the differentiating component because it shows different meanings between the words being contrasted.

4.5 Put a Proper Microstructure of the Dictionary

A dictionary of course must contain various information related to the entries included in it, including meaning, word categories or word classes, pronunciations, meanings, polysemes, homonyms, example sentences, and additional explanations relating to typical entries. Lexicographers must possess knowledge of microstructures in making a bilingual dictionary.

A term is a word, an expression or alphanumeric symbol used by expert in specialized technical subject to designate a concept. Term and concept which are unity is an essential requirement of unambiguous communications (Hartmann and James, 1998: 138-139 in Vrbinc).

KBJBI is not complete yet in microstructure. The most basic thing that is not listed in it is pronunciation, and this is actually fatal when it comes to making dialect dictionaries. Compared to the *Kamus Dialek Banyumas* (2007), KBJBI is arguably better. However, improvements need to be made so that the bilingual dictionary function can be maximized. Providing example sentences must also be made as natural as possible so that the function as a dialect dictionary for the general public will be maximized.

5 Conclusion

After examining the lexicographer strategy in making the KBJBI as lexicographic evidence of maintenance local wisdom, several conclusions can be drawn. The selection of entries in KBJBI still uses conventional methods, so it needs updating. Digitalization efforts are also needed for a more comprehensive language distribution, such as corpora base analysis. Information about the type of dictionary has not been well manifested, whether productive or receptive. There is no pronunciation which is important in a dialect dictionary. There are many obscurities, especially in synonym, so the meaning component analysis is needed to comprehended the next project. The entries listed in the KBJBI reflect local wisdom, namely the factual manifestation of Banyumasan people's life which includes culture, habits, and life systems.

The processes that have been carried out as a strategy for making dialect dictionaries, starting from data selection, making the proper definition, local cultural knowledge, understanding the sense of meaning, and good microstructural knowledge will create a dialect dictionary result that can reflect the local wisdom of the speaking community's culture. This requires continuous effort and learning.

6 References

Arnaud, Sabine. (2019). "From Gesture to Sign: Sign Language Dictionaries ang The Invention of A Language". Sign A Language Studies, Vol 20/1: 41-82. Gallaunt University Press. URL: https://www.jstor.org/stable/10.2307/26899343

Atkins, Sue and Rundell, Michael. (2008). *The Oxford Guide to Practical Lexicography.* United States: Oxfrod University Press.

Budiwiyanto, Adi. http://badanbahasa.kemdikbud.go.id/lamanbahasa/artikel/2796/kamus- dalam-perspektif-budaya-material

Chaer, Abdul. (2007). *Linguistik Umum.* Jakarta: Rineka Cipta

Crawfort, Hannah. (2014). "Linguistics, Lexicography, and The Early Modern". Journal for Early Modern Cultural Studies, Vol. 14, No.2, 2014: 94—99. University of Pennsylvania Press.

Kridalaksana, Harimurti. (2003). "Kamus Sebagai Dokumentasi Bahasa: Dalam Laporan Sanggar Kerja Internasional tentang Leksikologi: Rintisan dalam Kajian Leksikologi dan Leksikografi, 16—17 Desember 2002. Depok: Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia: 165—167.

Guralnik, David B. (1958). "Connotation in Dictionary Definition". College Composition and Communication, May 1958, Vol 9/2: 90-93. National Council of Teachers of English. URL: https://www.jstor.org/stable/355330

Hartmann, R.R.K. & Gregery James. (2002). *Dictionary of Lexicography.* London and New York: Routledge.

Henderson, John. (No Year). Assesing Carl Strehlow's Dictionary as Linguistic Description: Present Value and Future Potential. URL: https://www.jstor.org/stable/j.ctv5cgb2c.13

Moedjanto, G. (1987). *Konsep-Kekuasaan Jawa: Penerapannya oleh Raja-Raja Mataram.* Yogyakarta: Penerbit Kanisius.

Moerdijk, Fons. (2003). "The Codification of Semantic Information". *Practical Guide to Lexicography* edited by Piet Van Sterkenburg. Amsterdam: John Benjamins Publishing Company

Nida, Eugene A. (1975). *Componential Analysis of Meaning.* Netherlands: Mouton & Co. N.V., Publishers, The Hague 4 Use APA6 for the documentation. List only the 5 major references.

Paryono, Yani. (2011). "Keunikan Bahasa Jawa Dialek Banyumas Sebagai Cerminan Identitas Masyarakat Banyumas". Makalah disampaikan pada Kongres Bahasa Jawa V di Surabaya tanggal 27—30 November 2011

Putra, Heddy Shri Ahimsa. (2009). "Bahasa, Sastra, dan Kearifan Lokal di Indonesia". *Mabasan* Vol 3/1: 30-57

Sterkenburg, Piet Van. 2003. *A Practical Guide to Lexicography.* Amsterdam: John Benjamins Publishing Company

Tohari, Ahmad, M. Koderi. (2007). *Kamus Dialek Banyumas.* Banyumas: Yayasan Carablaka

Vrbinc, Marjeta & Alenka Vrbinc. "Subject-Field in Monolingual Learners Dictionaries: The Gap between The Current State and Dictionary User's Expectations. URL: https://www.jstor.org/stable/10.2307/26430939

Wahyuni, Tri. (2014). "Kajian Perbandingan Bahasa Jawa Standar dengan Bahasa Jawa Banyumasan". Prosiding Seminar Internasional Kajian Leksikologi dan Leksikografi Mutakhir "Pelbagai Persoalan Penyusunan Kamus dan Pelaksanaan Undang-Undang Bahasa RI di Ranah Publik, khususnya di dalam Leksikologi dan Leksikografi: 199--214. Depok: Laboratorium Leksikografi, FIB, Universitas Indonesia. https://id.scribd.com/doc/288230555/Pros

Wahyuni, Tri et al. (2015). *Kamus Bahasa Jawa Banyumasan-Indonesia.* Semarang: Balai Bahasa Provinsi Jawa Tengah

DIGITALIZING LOCAL LANGUAGE DICTIONARY: CHALLENGES AND OPPORTUNITIES

Winda Luthfita, Selly Rizki Yanita
National Agency for Language Development and Cultivation, Indonesia
winda.luthfita@kemdikbud.go.id; selly.rizki@kemdikbud.go.id

Abstract

Badan Pengembangan dan Pembinaan Bahasa (National Agency of Language Development and Cultivation), as a government agency under The Ministry of Education and Culture of Indonesia, has published many dictionaries. There are more than 100 dictionaries have been published since 1977. Some dictionaries have been revised by adding new entries and senses. With the massive technology development, people also change the way they live and how they see life. Everything needs to be accessible to everyone, anytime and anywhere, lightly and easily. To contribute to this new technology era, Badan Pengembangan dan Pembinaan Bahasa has started the integration project that aims to provide the online application of the language products of Badan Pengembangan dan Pembinaan Bahasa. It has started by launching The Program Pengayaan Kosakata (word proposal application in 2015) and was followed by launching the online version of Kamus Besar Bahasa Indonesia (The official dictionary of the Indonesian Language), Tesaurus Tematis Bahasa Indonesia, and Ensiklopedia Sastra Indonesia in 2016. In 2020, Badan Pengembangan dan Pembinaan Bahasa has started the development of Aplikasi Pangkalan Data Kamus or also called Aplikasi Kompilasi Kamus. This online application will accommodate at least three kinds of dictionaries: the local language dictionary, technical term dictionary, and bilingual dictionary published by Badan Pengembangan dan Pembinaan Bahasa. The process is continued by developing a digitalization project targeting the digitalization of printed versions of specific term dictionaries, Indonesian-local language dictionaries, and local language-Indonesian dictionaries. This paper aims to provide the process of digitalization, challenges, and opportunities that come along the process. The research method uses qualitative methods of observing the file of dictionaries and collecting some challenges through the whole process. The results of this study are expected to support the digitalization process and dictionary development in Indonesia.

Keywords Local language, digitalization, online dictionary

1 Introduction

National Agency of Language Development and Cultivation is a government organization under The Ministry of Education and Culture of The Republic of Indonesia responsible for language development. The massive development of technology changes every aspect of life, including language information needs. According to that, the Agency is continuously improving lexicography products, including doing digitalization. Digitalization is one of the lexicography programs that aim to provide a digital version of printed dictionaries that have been published by the National Agency of Language Development and Cultivation. To supporting that, the database of dictionary and application is made. Three kinds of dictionaries are digitalized: local language-Indonesian dictionary, Indonesian-local language dictionary, and technical term dictionary. Besides providing easily accessed word information on various languages or specific science, digitalization of dictionaries program is expected to be the better alternative of word documentation, providing various alternative word translations, and supporting Kamus Besar Bahasa Indonesia (abbreviated as KBBI). There are more than 100 dictionaries that are expected to be digitalized.

The program will be conducted for five years, which is classified by this timetable.

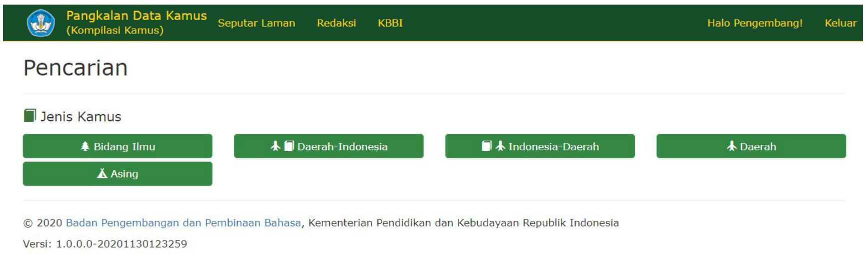
Table 1 Time table of Digitalizing Dictionary Project

Year	Description
2020	1) Application developing 2) Printed dictionary inventarization
2021	1) Data scanning 2) Data selection 3) Data input - Data input to template and application - Data verification 4) Application updating
2022	1) Data input - Data input to template and application - Data verification 2) Application updating
2023	1) Data input - Data input to template and application - Data verification 2) Application updating
2024	1) Data input - Data input to template and application - Data verification 2) Application updating

2 Application and Lexicographical Work

2.1 Application developing

When we talk about digitalization, it is essential to know how the infrastructure will be developed. In this regard, The Agency tried to develop a non-monolithic application. Non-monolithic application is an application that has several features and can accommodate several product databases. The Agency learned that non-monolithic application is the best option in terms of budget and time efficiency. The Agency started the application development in 2020 and is called a dictionary compilation application. The application is expected to document (compile, develop, and edit) four kinds of dictionaries, including local language-Indonesian dictionary, Indonesian-local language dictionary, technical term dictionary, and foreign language dictionary. It is designed to archive different words from the local languages of Indonesia, which the representative offices of The National Agency continuously collect. The archive system is integrated into KBBI website.



© 2020 Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan Republik Indonesia

Versi: 1.0.0.0-20201130123259

Figure 1 Dictionary Compilation Application

2.2 Application architecture

As previously mentioned, the application is designed based on the lexicography work process of the National Agency. The application has an editor page classified by five kinds of sections: raw proposal, editor desk (classified by the dictionary category), word list, admin, and dictionary update section.

All of the digitalized dictionaries will be proposed by data template on "raw proposal section". The template is a specially formulated excel that can be integrated into the application. After the dictionary information is put on the excel template, the template will be uploaded on "upload subsection", and the uploaded files will be found on the "list subsection". In contrast, the archived files can be found on the "archive subsection".

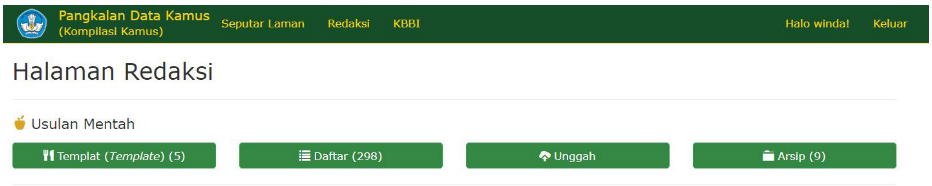


Figure 2 Raw Proposal Section

After the data template has been uploaded and getting precise test results, the data will be forwarded to the editor desk. The validator will check whether the data meet the lexicographic standard or not. The checking process can be conducted one by one or massively. The data that meets the standard will be validated and forwarded to "validated" subsection. The coordinator will check the validated proposal and conduct a dictionary update. The proposal will be forwarded to "accepted" subsection and can be searched in "searching" section. As the application is integrated into KBBI application, the data that is available to search on Dictionary Compilation Application can be the base word proposal of KBBI. Finally, the data on Dictionary Compilation Application can be the supporting system of the KBBI update program.

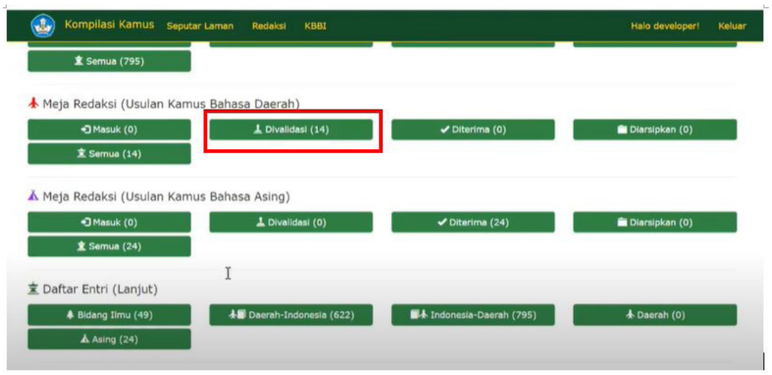
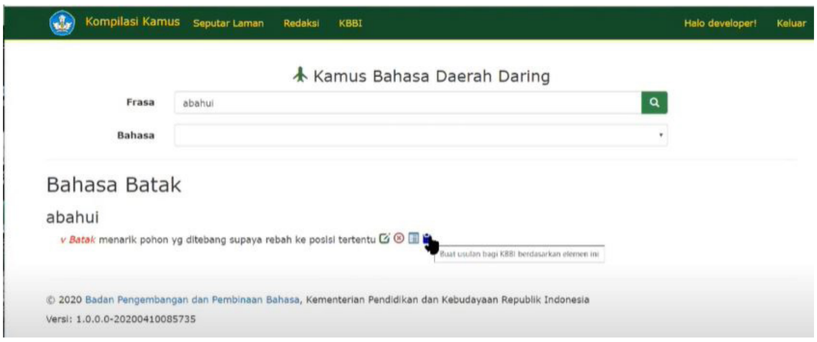


Figure 3 Editor Desk and Validation Subsection



Kamus Bahasa Daerah Daring

Frasa: abahui

Bahasa:

Bahasa Batak

abahui

v Batak menarik pohon yg ditebang supaya rebah ke posisi tertentu

Buat usulan bagi KBBI berdasarkan elemen ini

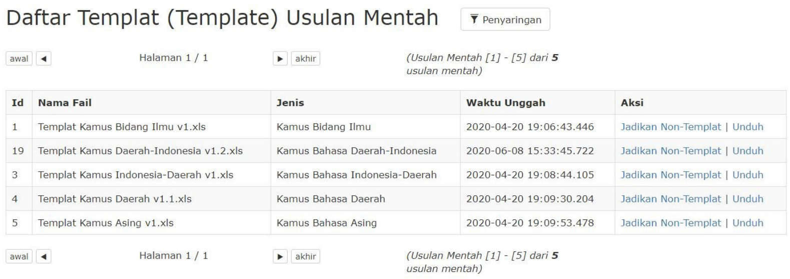
© 2020 Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan Republik Indonesia

Versi: 1.0.0.0-20200410085735

Figure 4 Searching Section and KBBI Proposal Feature

2.3 Data input

The Data input process was started in 2020. It is a part of the trial and error process of the data template and application testing stage. Five kinds of data templates can be used to make word proposals to the application, namely technical term dictionary template, local language-Indonesia dictionary template, Indonesia-local language dictionary template, local language dictionary template, and foreign language dictionary template. As previously mentioned, the template is a particularly formulated excel that contains specific input rules. All staff involved in any given level got a technical briefing from the developer to guarantee the data quality.



Daftar Templat (Template) Usulan Mentah

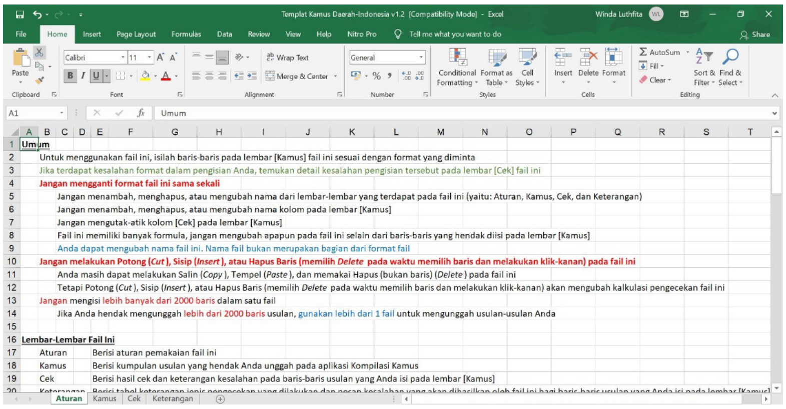
(Usulan Mentah [1] - [5] dari ***5*** *usulan mentah)*

Id	Nama Fail	Jenis	Waktu Unggah	Aksi
1	Templat Kamus Bidang Ilmu v1.xls	Kamus Bidang Ilmu	2020-04-20 19:06:43.446	Jadikan Non-Templat | Unduh
19	Templat Kamus Daerah-Indonesia v1.2.xls	Kamus Bahasa Daerah-Indonesia	2020-06-08 15:33:45.722	Jadikan Non-Templat | Unduh
3	Templat Kamus Indonesia-Daerah v1.xls	Kamus Bahasa Indonesia-Daerah	2020-04-20 19:08:44.105	Jadikan Non-Templat | Unduh
4	Templat Kamus Daerah v1.1.xls	Kamus Bahasa Daerah	2020-04-20 19:09:30.204	Jadikan Non-Templat | Unduh
5	Templat Kamus Asing v1.xls	Kamus Bahasa Asing	2020-04-20 19:09:53.478	Jadikan Non-Templat | Unduh

(Usulan Mentah [1] - [5] dari ***5*** *usulan mentah)*

Figure 5 The Data Template

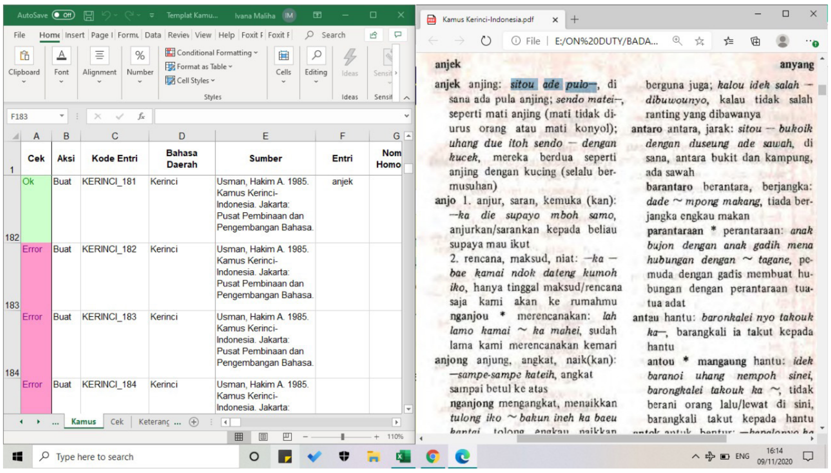
Every template has four sheets: rule, dictionary, check, and note. The rule sheet explained basic rules and instructions of template use; the dictionary sheet has columns for input the lemma information that a dictionary has. The check sheet has several columns that describe whether the data on the dictionary sheet is correct or wrong. The incorrect data will be classified as an 'error'. Last but not least is the note sheet. This sheet explains actions that can be done to the template and the relevant note. It also counts the status of data input classified to 'ok' or 'error' column.



1 **Umum**
2 Untuk menggunakan fail ini, isilah baris-baris pada lembar [Kamus] fail ini sesuai dengan format yang diminta
3 Jika terdapat kesalahan format dalam pengisian Anda, temukan detail kesalahan pengisian tersebut pada lembar [Cek] fail ini
4 **Jangan mengganti format fail ini sama sekali**
5 Jangan menambah, menghapus, atau mengubah nama dari lembar-lembar yang terdapat pada fail ini (yaitu: Aturan, Kamus, Cek, dan Keterangan)
6 Jangan menambah, menghapus, atau mengubah nama kolom pada lembar [Kamus]
7 Jangan mengutak-atik kolom [Cek] pada lembar [Kamus]
8 Fail ini memiliki banyak formula, jangan mengubah apapun pada fail ini selain dari baris-baris yang hendak diisi pada lembar [Kamus]
9 Anda dapat mengubah nama fail ini. Nama fail bukan merupakan bagian dari format fail
10 **Jangan melakukan Potong (*Cut*), Sisip (*Insert*), atau Hapus Baris (memilih *Delete* pada waktu memilih baris dan melakukan klik-kanan) pada fail ini**
11 Anda masih dapat melakukan Salin (*Copy*), Tempel (*Paste*), dan memakai Hapus (bukan baris) (*Delete*) pada fail ini
12 Tetapi Potong (*Cut*), Sisip (*Insert*), atau Hapus Baris (memilih *Delete* pada waktu memilih baris dan melakukan klik-kanan) akan mengubah kalkulasi pengecekan fail ini
13 Jangan mengisi lebih banyak dari 2000 baris dalam satu fail
14 Jika Anda hendak mengunggah lebih dari 2000 baris usulan, gunakan lebih dari 1 fail untuk mengunggah usulan-usulan Anda
15
16 **Lembar-Lembar Fail Ini**
17 Aturan Berisi aturan pemakaian fail ini
18 Kamus Berisi kumpulan usulan yang hendak Anda unggah pada aplikasi Kompilasi Kamus
19 Cek Berisi hasil cek dan keterangan kesalahan pada baris-baris usulan yang Anda isi pada lembar [Kamus]
20 Keterangan Berisi tabel keterangan jenis pengecekan yang dilakukan dan pesan kesalahan yang akan dihasilkan oleh fail ini bagi baris-baris usulan yang Anda isi pada lembar [Kamus]

Figure 6 Rule Sheet

Before data input was started, The Language Agency conducted data selection by collecting all published dictionaries from library repository or representative offices. We found two types of dictionaries that The Language Agency has, including pdf and printed dictionary. The printed dictionary needs to be scanned before being input into the template. The data staff has two options for inputting data to the template: retype or copy-paste the pdf data to every column on a template or convert the pdf type to word file type and copy-paste the data after that.



	A	B	C	D	E	F	G
1	Cek	Aksi	Kode Entri	Bahasa Daerah	Sumber	Entri	Nom Homo
182	Ok	Buat	KERINCI_181	Kerinci	Usman, Hakim A. 1985. Kamus Kerinci-Indonesia. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.	anjek	
183	Error	Buat	KERINCI_182	Kerinci	Usman, Hakim A. 1985. Kamus Kerinci-Indonesia. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.		
184	Error	Buat	KERINCI_183	Kerinci	Usman, Hakim A. 1985. Kamus Kerinci-Indonesia. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.		
	Error	Buat	KERINCI_184	Kerinci	Usman, Hakim A. 1985. Kamus Kerinci-Indonesia. Jakarta:		

anjek **anyang**

anjek anjing: *sitou ade pulo—*, di sana ada pula anjing; *sendo matei—*, seperti mati anjing (mati tidak diurus orang atau mati konyol); *uhang due itoh sendo — dengan kucek*, mereka berdua seperti anjing dengan kucing (selalu bermusuhan)

anjo 1. anjur, saran, kemuka (kan): *—ka die supayo mboh samo*, anjurkan/sarankan kepada beliau supaya mau ikut
2. rencana, maksud, niat: *—ka — bae kamai ndok dateng kumoh iko*, hanya tinggal maksud/rencana saja kami akan ke rumahmu
nganjou * merencanakan: *lah lamo kamai ~ ka mahei*, sudah lama kami merencanakan kemari

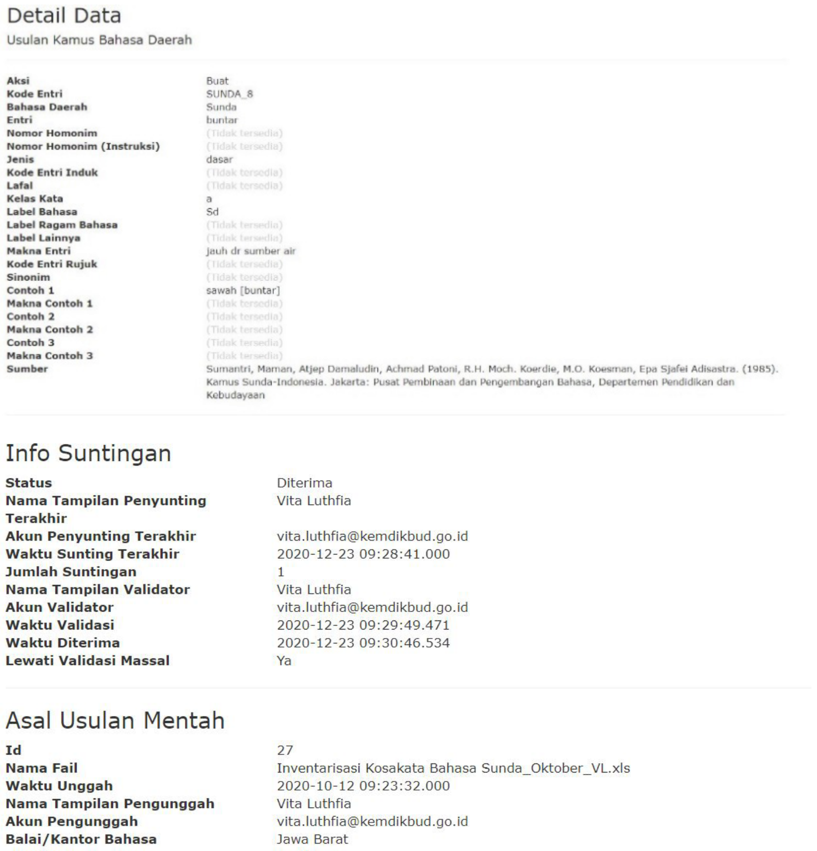
anjong anjung, angkat, naik(kan): *—sampe-sampe kateih*, angkat sampai betul ke atas
nganjong mengangkat, menaikkan *tulong iko ~ bakun ineh ka baeu bantai* tolong engkau naikkan

berguna juga; *kalou idek salah — dibuwounyo*, kalau tidak salah ranting yang dibawanya

antaro antara, jarak: *sitou — bukoik dengan duseung ade sawah*, di sana, antara bukit dan kampung, ada sawah
barantaro berantara, berjangka: *dade ~ mpong makang*, tiada berjangka engkau makan
parantaraan * perantaraan: *anak bujon dengan anak gadih mena hubungan dengan ~ tagane*, pemuda dengan gadis membuat hubungan dengan perantaraan tua-tua adat

antau hantu: *baronkalei nyo takouk ka—*, barangkali ia takut kepada hantu
antou * mangaung hantu: *idek baranoi uhang nempoh sinei, barongkalei takouk ka ~*, tidak berani orang lalu/lewat di sini, barangkali takut kepada hantu

Figure 7 Data Input Process



Detail Data

Usulan Kamus Bahasa Daerah

Aksi	Buat
Kode Entri	SUNDA_8
Bahasa Daerah	Sunda
Entri	buntar
Nomor Homonim	(Tidak tersedia)
Nomor Homonim (Instruksi)	(Tidak tersedia)
Jenis	dasar
Kode Entri Induk	(Tidak tersedia)
Lafal	(Tidak tersedia)
Kelas Kata	a
Label Bahasa	Sd
Label Ragam Bahasa	(Tidak tersedia)
Label Lainnya	(Tidak tersedia)
Makna Entri	jauh dr sumber air
Kode Entri Rujuk	(Tidak tersedia)
Sinonim	(Tidak tersedia)
Contoh 1	sawah [buntar]
Makna Contoh 1	(Tidak tersedia)
Contoh 2	(Tidak tersedia)
Makna Contoh 2	(Tidak tersedia)
Contoh 3	(Tidak tersedia)
Makna Contoh 3	(Tidak tersedia)
Sumber	Sumantri, Maman, Atjep Damaludin, Achmad Patoni, R.H. Moch. Koerdie, M.O. Koesman, Epa Sjafei Adisastra. (1985). Kamus Sunda-Indonesia. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan

Info Suntingan

Status	Diterima
Nama Tampilan Penyunting Terakhir	Vita Luthfia
Akun Penyunting Terakhir	vita.luthfia@kemdikbud.go.id
Waktu Sunting Terakhir	2020-12-23 09:28:41.000
Jumlah Suntingan	1
Nama Tampilan Validator	Vita Luthfia
Akun Validator	vita.luthfia@kemdikbud.go.id
Waktu Validasi	2020-12-23 09:29:49.471
Waktu Diterima	2020-12-23 09:30:46.534
Lewati Validasi Massal	Ya

Asal Usulan Mentah

Id	27
Nama Fail	Inventarisasi Kosakata Bahasa Sunda_Oktober_VL.xls
Waktu Unggah	2020-10-12 09:23:32.000
Nama Tampilan Pengunggah	Vita Luthfia
Akun Pengunggah	vita.luthfia@kemdikbud.go.id
Balai/Kantor Bahasa	Jawa Barat

Figure 8 Word Proposal

Nine dictionaries have been proposed or still in process for template proposal as a part of the application testing stage, including four technical term dictionaries, two Indonesian-local language dictionaries, and three local language-Indonesian dictionaries. As a part of the testing stage, we try to find various dictionary characteristics to find out whether the template can accommodate the word information of the dictionary. Some situations have been noted and will be delivered more in the analysis section of this paper. The data input for the local language dictionary will be continued in June 2021.

2.4 Lexicographical workflow

Besides making the better word documentation system, the application is made to make the lexicographic work easier. Designing based on the lexicography work process in the Agency, the application has a multilevel inspection feature that makes the word template checked and corrected a few times by the editor, validator, and coordinator. Each level has its authority which is described as follows.

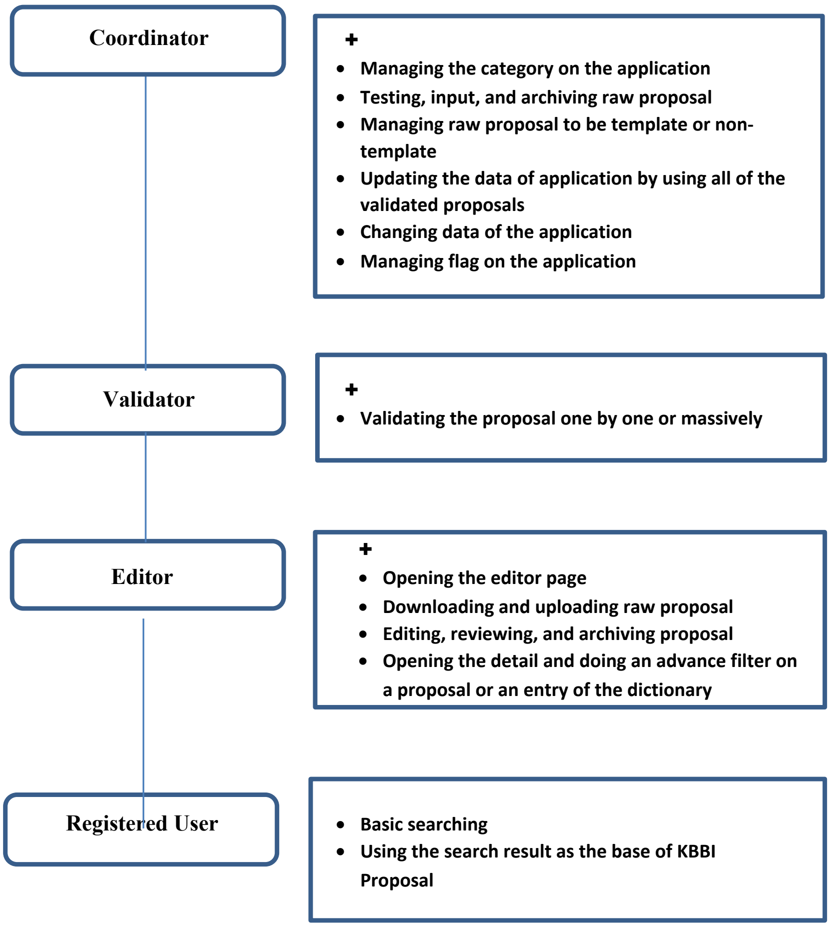

Coordinator
+
• Managing the category on the application
• Testing, input, and archiving raw proposal
• Managing raw proposal to be template or non-template
• Updating the data of application by using all of the validated proposals
• Changing data of the application
• Managing flag on the application
Validator
+
• Validating the proposal one by one or massively
Editor
+
• Opening the editor page
• Downloading and uploading raw proposal
• Editing, reviewing, and archiving proposal
• Opening the detail and doing an advance filter on a proposal or an entry of the dictionary
Registered User
• Basic searching
• Using the search result as the base of KBBI Proposal

Figure 9 The Lexicographical Workflow

Currently, the application can be accessed by specific users granted the right to access and is used for internal purposes only as the development is still in process. The application has several features, which are described as follows.

1. Multi-dictionaries searching
2. Massive word proposal and acceptance
3. Limited grant access

3 Challenges in Digitalizing

The implementation of digitalizing dictionaries faces several obstacles due to several factors, including limited human resources, the condition of the dictionary itself, and differences in the microstructure format of local language dictionaries. These three factors are detailed as follows.

3.1 Human Resources

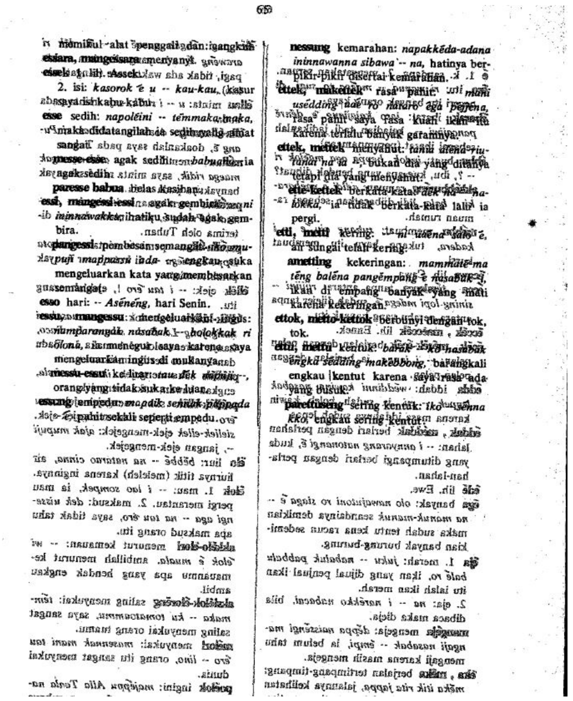
In planning for digitalization, the aspect of human resources is essential to be determined. Hughes (2004, p. 96–97) states that there are two domains of resources that need to be considered in planning: resources involved in digitization activities and resources that will maintain the digitalization. In the digitalization process, we need a lot of human resources. It is because the work is not only inputting data but also collecting, checking, rechecking, and even adding new data information.

Dictionary building is a continuous program in the representative offices all around Indonesia, well known as Kantor Bahasa and Balai Bahasa, every year. Therefore, there are so many local language dictionaries that need to be digitalized. The number of dictionaries to be inputted into the application is more than 100 dictionaries. Unfortunately, the limited number of human resources made us recruit outsourcing workers to input data. The data complexity requires specific lexicography competency, patience, and also determination. Hence, the hiring process needs more attention as the data input process is quite tricky. Indeed, the existence of data input staff makes the work easier. However, the verification process, including checking and rechecking data, is another thing to be considered. As the number of lexicographers on our Agency is limited, the data processing will be challenging. At the same time, the lexicographer also needs to understand the whole business process and perform administrative work regarding outsourcing hiring. Understanding the whole business process and every detail is required as the data input process hires outsourcing workers who do not work permanently. Therefore, the training process needs to be done every time the outsourcing worker changes.

On top of that, the lexicographers also have another program to do. Therefore, the time management and work distribution are pretty challenging as the lexicographer needs to ensure that the data meet the requirement for the expected use. Despite the data input worker, the staff competency is the area we try to work for. Our Agency needs application developers, even a team of developers, who permanently work as staff in our Agency. Our Agency now hires an independent developer to develop, update, and maintain the application. If the application will be regularly developed, the team responsible for working on the database, application system, security system, and user interface design is more than needed.

3.2 Condition of Dictionary

As mentioned above, almost every year, the local language dictionary produced. It makes the abundance of local language dictionary products. Unfortunately, the storage of these dictionaries is not centralized in one place. Usually, they send dictionaries in print to the library, not in digital (PDF format). It is difficult for us to track and collect the dictionary, especially the earlier and old dictionaries. We also find out that the earlier and old dictionaries available in specific years only. More investigation is needed to find out the dictionary development program conducted in that year and the existence of the dictionaries. The available dictionary data process is also challenging. As the digital format is not available, we need to scan the printed dictionary after tearing the cover and binding. This activity is pretty time-consuming. In addition, the physical condition of the dictionary whose paper has been eaten by termites, torn, crumpled, and illegible writing due to the old age of the dictionary makes in poor-quality scans so that the information in the dictionary can be incomplete. Some dictionaries with those kinds of conditions are available on pdf only. Therefore, extra effort is needed to find the physical book and make sure the information on the dictionary.



65

ri memikul alat penggali dan angkat
essara, mangessara menyanyi.
essek lih. **Assek**.
2. isi: *kasorok e u -- kau-kau*, (kasur saya isinya kabu-kabu.
esse sedih: *napoléini -- tèmmaka-maka*, maka didatangilah ia sedih yang amat sangat.
messe-esse agak sedih: *babuana* ia agak sedih.
paresse babua belas kasihan.
essi, mangessi-essi agak gembira: ...*ni innawakku*, hatiku sudah agak gembira.
pangessi pembesar semangat: *iko mupuji mappaessi iada-* engkau suka mengeluarkan kata yang membesarkan semangat.
esso hari: -- *Aséneng*, hari Senin.
essu, mangessu mengeluarkan ingus: *nlumparangdk nasabak -- bolokkak ri olona*, saya menegur karena saya mengeluarkan ingus di mukanya.
messu-essu keluar: *tau dék maelo* --, orang yang tidak suka keluar.
essung empedu: *mapaik sennak pappada* --, pahit sekali seperti empedu.

nessung kemarahan: *napakkéda-adana ininnawanna sibawa -- na*, hatinya berpikir-pikir disertai kemarahan.
ettek, makettek rasa pahit: *mani usédding uae ro nakaro ega pejjena*, rasa pahit saya rasa kuah ikan itu karena terlalu banyak garamnya.
ettek, mettek menyahut: *tania alena riutanai na ia* --, bukan dia yang ditanya tetapi dia yang menyahut.
ette-kettek berkata-kata: *dék na -- na-lokka*, ia tidak berkata-kata lalu ia pergi.
etti, metti kering: -- *ni uaena salo e*, air sungai telah kering.
ametting kekeringan: *mammate ma téng baléna pangémpang e nasabak -- i*, ikan di empang banyak yang mati karena kekeringan.
ettok, metto-kettok berbunyi dengan tok, tok.
ettu, mettu kentut: *barak -- ko nasabak engka sedding makebbong*, barangkali engkau kentut karena saya rasa ada yang busuk.
parettuseng sering kentut: *iko -- senna kko*, engkau sering kentut.

Figure 10 Bugis-Indonesia Dictionary File Condition

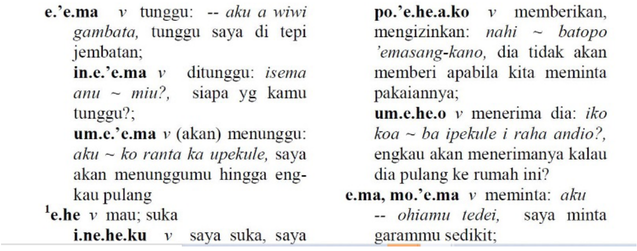
3.3 Microstructure Format

The third challenge that we found when inputting data is dictionary format, especially in microstructures. Microstructure referred to the internal composition of dictionaries. Microstructures consist of headword, spelling, pronunciation, usage label, part of speech, definition, phrase or sentence example, etymology, cross-reference, and semantic domain (Rehg, 2018). The lexicographers have their styles when designing the microstructure format. It causes diversity in the microstructure writing format. Here are the examples.

3.3.1 Headword

The headword shows how lemmas are written, whether in a single word, a hyphenated word, or several words, and some dictionaries show with the wordbreaks (Atkins and Rundell, 2008). Every dictionary has its own convention of the presentation style. For example, the headword of Online KBBI is written with wordbreaks, which referred to Pedoman Umum Ejaan Bahasa Indonesia (General Guidelines for Indonesian Spelling), while Merriam-Webster Online Dictionary and Oxford Online Dictionary use the word without wordbreaks style.

Differences in the writing style of headwords are also found in some printed local language dictionaries. For example, in Mori-Indonesia Dictionary and Mbojo-Indonesia Dictionary, the words are written with wordbreaks, while the other dictionaries without wordbreaks. The use of wordbreaks, especially in local language dictionaries, can make it easier for dictionary users to spell the word. In addition to wordbreaks, headwords are written with a phonetic symbol or diacritical marks for certain letters. It is used to facilitate the pronunciation of words. The diacritical marks commonly used for dictionaries are accent marks, such as acute, grave, and circumflex. It indicates the different types of pitch accents for letter e (e.g., *aé*-water and *kênal*-know). Another mark is a macron line to mark long or heavy letters. However, we found that in The Alas-Indonesia dictionary, the macron line is used to symbolize the letter e that sounds [é] and [è].



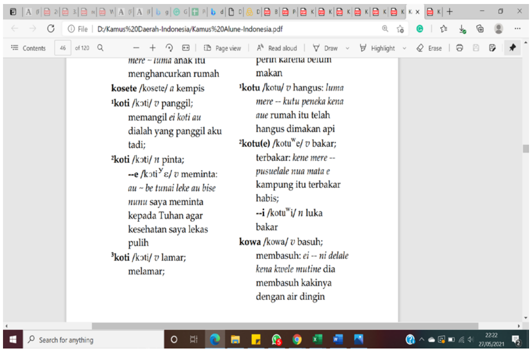
e.'e.ma *v* tunggu: -- *aku a wiwi gambata,* tunggu saya di tepi jembatan;
in.e.'e.ma *v* ditunggu: *isema anu ~ miu?,* siapa yg kamu tunggu?;
um.e.'e.ma *v* (akan) menunggu: *aku ~ ko ranta ka upekule,* saya akan menunggumu hingga engkau pulang
^{1}**e.he** *v* mau; suka
i.ne.he.ku *v* saya suka, saya
po.'e.he.a.ko *v* memberikan, mengizinkan: *nahi ~ batopo 'emasang-kano,* dia tidak akan memberi apabila kita meminta pakaiannya;
um.e.he.o *v* menerima dia: *iko koa ~ ba ipekule i raha andio?,* engkau akan menerimanya kalau dia pulang ke rumah ini?
e.ma, mo.'e.ma *v* meminta: *aku -- ohiamu tedei,* saya minta garammu sedikit;

Figure 11 Example of Headword in Mori-Indonesia Dictionary

3.3.2 Pronunciation

Pronunciation is an essential element, especially in the bilingual dictionary. It helps the users to pronounce the lemma correctly. According to Klapicová (2005), when preparing the grammatical indication of a bilingual dictionary, the lexicographer should consider that they are written for foreigners, not for native speakers. Thus, the lemma should indicate the pronunciation of the entry in its canonical form.

From 13 local language-Indonesia dictionaries, we found that only five dictionaries provided pronunciation, most of them only for letter e. However, those dictionaries have a different style in presenting the pronunciation. In Talaud-Indonesia Dictionary and Alune-Indonesia Dictionary, pronunciation is written with slashes in phonemic transcription (e.g., *kohu*-wave /kohu/) while pronunciation in Angkola Mandailing-Indonesia is written with brackets in phonetic transcription (e.g., *dayuk*-soft [dayu:k]). The different styles of it can be a problem when we input it into the application. For example, The Alune-Indonesia dictionary wrote semivowel sounds in pronunciation (e.g., *kuala*-flirt [ku^{w}ala]), an uncommon form in dictionary pronunciation. Therefore, the lexicographer should equate the data that take much time.



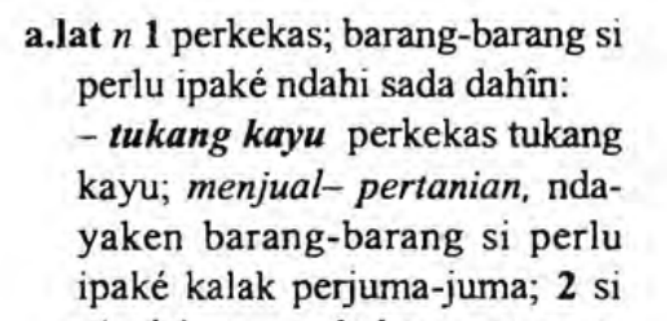
mere ~ luma anak itu menghancurkan rumah
kosete /kosete/ *a* kempis
^{1}**koti** /kɔti/ *v* panggil;
memangil *ei koti au* dialah yang panggil aku tadi;
^{2}**koti** /kɔti/ *n* pinta;
--e /kɔti^{y}ε/ *v* meminta: *au ~ be tunai leke au bise nunu* saya meminta kepada Tuhan agar kesehatan saya lekas pulih
^{3}**koti** /kɔti/ *v* lamar;
melamar;
perih karena belum makan
^{1}**kotu** /kotu/ *v* hangus: *luma mere -- kutu peneka kena aue* rumah itu telah hangus dimakan api
^{2}**kotu(e)** /kotu^{w}e/ *v* bakar;
terbakar: *kene mere -- pusuelale nua mata e* kampung itu terbakar habis;
--i /kotu^{w}i/ *n* luka bakar
kowa /kowa/ *v* basuh;
membasuh: *ei -- ni delale kena kwele mutine* dia membasuh kakinya dengan air dingin

Figure 12 Pronunciation Style in Alune-Indonesia Dictionary

3.3.3 Equivalent

The definition is an essential element in the monolingual dictionary macrostructure. Every user needs the meaning of the word definitively. It is also applied in bilingual dictionaries. The user of a bilingual dictionary wants an exact equivalent of terms which appropriate to the context. Otherwise, lexicographers should have to provide the equivalent of terms precisely in dictionaries (Wojowasito, 2007). In the local language dictionary that we used, most of them provide the equivalent of terms as close as possible.

Nonetheless, we found that in Indonesia-Karo Dictionary, there are two ways to present the equivalent. First, it presents with equivalent or synonym of terms in the Karo language. Second, it was given by an explanation such as definition in monolingual dictionary and continued by a synonym. The second way is not efficient as an equivalent as Newmark (1988) says that the translation of terms should be to achieve 'equivalent effect' which produce the same on the readership of the translation as was obtained on the readership of the original. Thus, lexicographers are forced to synchronize by excluding the explanation.

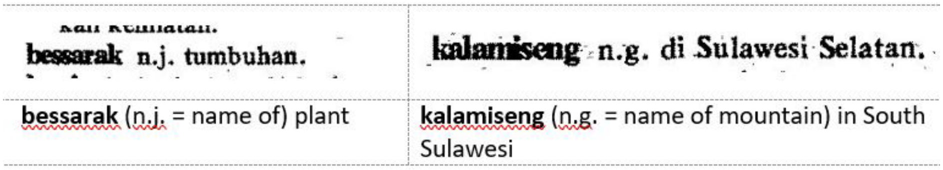


a.lat *n* **1** perkekas; barang-barang si perlu ipaké ndahi sada dahîn: – ***tukang kayu*** perkekas tukang kayu; ***menjual– pertanian,*** ndayaken barang-barang si perlu ipaké kalak perjuma-juma; **2** si

Figure 13 The Definition of Indonesia-Karo Dictionary

3.3.4 Usage Label

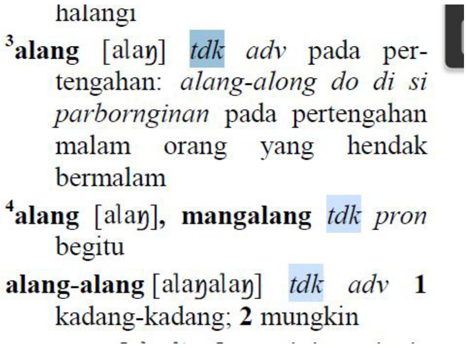
Usage label denotes the register, subject area, or other specific application of a word or phrase (Lexico, n.d.). It provides specific information about the domain of application of the definition. It restricts the definition to a certain context (Janssen, 2003). We found out that some labels are unique or not commonly used on most of our dictionaries based on our digitalization process. For instance, *n.j.* (abbreviation for *nama jenis*-name of kind of) used for classifying certain things, such as fish, fruit, animal, or plant and *n.g.* for *nama gunung*-mountain name to show the proper name of a mountain in certain areas on Bugis-Indonesia Dictionary.



bessarak n.j. tumbuhan.	**kalamiseng** n.g. di Sulawesi Selatan.
bessarak (n.j. = name of) plant	kalamiseng (n.g. = name of mountain) in South Sulawesi

Figure 14 n.j and n.g. usage label on Bugis-Indonesia Dictionary

This situation is also found in The Bali Kuno-Indonesia Dictionary that uses *n.* for showing *name of*. In Angkola Mandailing-Indonesia Dictionary, *tdk* (abbreviation for *tidak diketahui*-unknown') is used as a marker that native speakers are not familiar with the lemma. There is a high possibility that the word is ancient or a part of cultural or literature terminology.



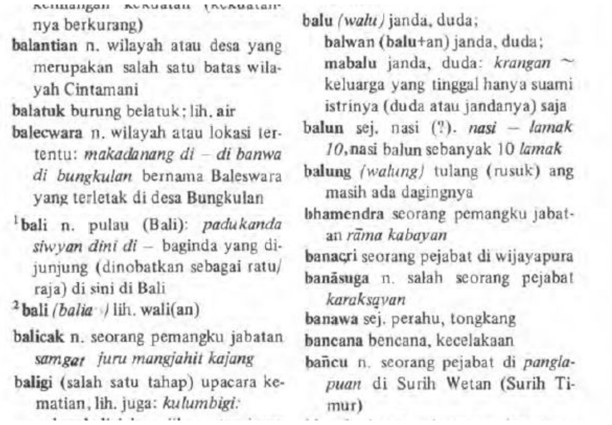
halangi

^{3}**alang** [alaŋ] *tdk adv* pada pertengahan: *alang-along do di si parbornginan* pada pertengahan malam orang yang hendak bermalam

^{4}**alang** [alaŋ], **mangalang** *tdk pron* begitu

alang-alang [alaŋalaŋ] *tdk adv* **1** kadang-kadang; **2** mungkin

Figure 15 n.j and n.g. usage label on Bugis-Indonesia Dictionary



kehilangan kekuatan (kekuatan-
nya berkurang)

balantian n. wilayah atau desa yang merupakan salah satu batas wilayah Cintamani

balatuk burung belatuk; lih. **air**

balecwara n. wilayah atau lokasi tertentu: *makadanang di – di banwa di bungkulan* bernama Baleswara yang terletak di desa Bungkulan

[1]**bali** n. pulau (Bali): *padukanda siwyan dini di –* baginda yang dijunjung (dinobatkan sebagai ratu/raja) di sini di Bali

[2]**bali** *(balia)* lih. **wali(an)**

balicak n. seorang pemangku jabatan *samgat juru mangjahit kajang*

baligi (salah satu tahap) upacara kematian, lih. juga: *kulumbigi:*

balu *(walu)* janda, duda;

balwan (**balu**+an) janda, duda;

mabalu janda, duda: *krangan ~* keluarga yang tinggal hanya suami istrinya (duda atau jandanya) saja

balun sej. nasi (?). *nasi – lamak 10*, nasi balun sebanyak 10 *lamak*

balung *(walung)* tulang (rusuk) ang masih ada dagingnya

bhamendra seorang pemangku jabatan *rāma kabayan*

banaçri seorang pejabat di wijayapura

banāsuga n. salah seorang pejabat *karaksạvan*

banawa sej. perahu, tongkang

bancana bencana, kecelakaan

bañcu n. seorang pejabat di *panglapuan* di Surih Wetan (Surih Timur)

Figure 16 n. Label for Indicating Geographic Location

(Region, Village, Island) and Position in Bali Kuno-Indonesia Dictionary

This condition was a challenge as the template does not accommodate this kind of case. In addition, more literature research is needed before inputting this data to the application to make sure the data. We also found out some abbreviations that do not commonly use in dictionary compiling details as follow. From 13 dictionaries that we try to observe, there are 5 dictionaries that have unique or different abbreviations that can be found in KBBI or another dictionary compiling in Center of Language Development and Preservation.

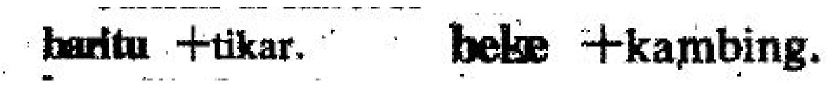
Table 2 List of Abbreviation

No.	Abbreviation	Meaning	Description	Dictionary	Year
1	adl	adalah	is	Jawa Tegal-Indonesia	2017
2	Adt	adat	tradition	Angkola Mandailing-Indonesia	2016
3	ag	agama	religion	Banjar-Indonesia	1977
4	Akl	Angkola	Angkola (language)	Angkola Mandailing-Indonesia	2016
5	bbrp	beberapa	several	Banjar Dialek Hulu-Indonesia	2008
6	bbs	bebasa (halus)	polite	Jawa Tegal-Indonesia	2017
7	bg	bagi	for	Banjar Dialek Hulu-Indonesia	2008
8	bg	bagian	part	Banjar-Indonesia	1977
9	Bh	bahasa halus	polite	Banjar-Indonesia	1977
10	Bk	bahasa kasar	harsh	Banjar-Indonesia	1977
11	blm	belum	not yet	Banjar Dialek Hulu-Indonesia	2008
12	bnt	binatang	animal	Banjar-Indonesia	1977
13	Bp	bahasa percakapan	informal (conversation)	Banjar-Indonesia	1977
14	dala	dl	in	Jawa Tegal-Indonesia	2017
15	dg	dengan	with	Banjar-Indonesia	1977
16	dl	dalam	in	Bali Kuno-Indonesia	1985
17	dl	dalam	in	Banjar-Indonesia	1977
18	dlm	dalam	in	Banjar Dialek Hulu-Indonesia	2008
19	dn	dengan	with	Bali Kuno-Indonesia	1985
20	dpt	dapat	can	Banjar-Indonesia	1977
21	kep	kependekan	abbreviation of	Alas-Indonesia	1985
22	kep	kependekan	abbreviation of	Banjar-Indonesia	1977
23	kp	kependekan	abbreviation of	Angkola Mandailing-Indonesia	2016

24	kt	kata	word	Banjar-Indonesia	1977
25	Mdl	Mandailing	Mandailing (language)	Angkola Mandailing-Indonesia	2016
26	nm	nama	name	Banjar-Indonesia	1977
27	ol	oleh	by	Jawa Tegal-Indonesia	2017
28	org	orang	person	Banjar-Indonesia	1977
29	peny	penyakit	disease	Banjar-Indonesia	1977
30	plg	paling	the most	Banjar Dialek Hulu-Indonesia	2008
31	sb	sebangsa	a kind of	Alas-Indonesia	1985
32	sblm	sebelum	before	Banjar Dialek Hulu-Indonesia	2008
33	sdg	sedang	while	Banjar Dialek Hulu-Indonesia	2008
34	sdh	sudah	after	Banjar Dialek Hulu-Indonesia	2008
35	sdh	sudah	has been	Banjar-Indonesia	1977
36	sdh	sudah	after	Jawa Tegal-Indonesia	2017
37	sej	sejenis	similar to …	Alas-Indonesia	1985
38	sej	sejenis	similar to …	Banjar Dialek Hulu-Indonesia	2008
39	sej	sejenis	similar to …	Banjar-Indonesia	1977
40	sf	sifat	character	Banjar-Indonesia	1977
41	shg	sehingga	so that	Banjar-Indonesia	1977
42	sj	sejenisnya	a kind of	Bali Kuno-Indonesia	1985
43	spy	supaya	in order to	Banjar-Indonesia	1977
44	ssdh	sesudah	after	Banjar Dialek Hulu-Indonesia	2008
45	sso	seseorang	someone	Jawa Tegal-Indonesia	2017
46	sst	sesuatu	something	Banjar Dialek Hulu-Indonesia	2008
47	sst	sesuatu	something	Jawa Tegal-Indonesia	2017
48	Tb	Toba	Toba (language)	Angkola Mandailing-Indonesia	2016
49	tdk	tidak dikenal	unknown	Angkola Mandailing-Indonesia	2016
50	tdk	tidak	not	Banjar-Indonesia	1977
51	tdk	tidak	not	Jawa Tegal-Indonesia	2017
52	tlh	telah	has done (something)	Banjar Dialek Hulu-Indonesia	2008
53	tlh	telah	has been	Banjar-Indonesia	1977
54	ttg	tentang	about	Banjar-Indonesia	1977
55	ttg	tentang	about	Jawa Tegal-Indonesia	2017
56	ttp	tetapi	but	Banjar Dialek Hulu-Indonesia	2008
57	ttp	tetapi	but	Banjar-Indonesia	1977
58	utk	untuk	for	Banjar Dialek Hulu-Indonesia	2008
59	ybs	untuk	for	Banjar-Indonesia	1977

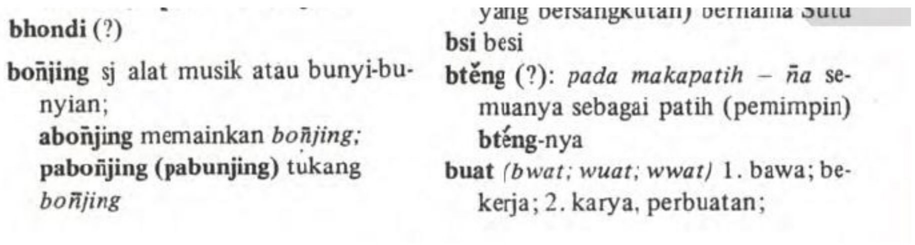
3.3.5 Another Component

Part of speech is a part of an essential component in a dictionary. Based on our digitalization process, we found that some dictionaries do not have any part of speech. There is also a particular symbol that indicates unique meaning, such as “?” in Bali Kuno-Indonesia dictionary to show a) in doubt, b) rarely used or only found once or twice on old Balinese inscription. There is also “+” on Bugis-Indonesia Dictionary to show that the word is rarely used in daily life (only found on Lontarak script), ancient word, or still in doubt.



baritu +tikar. **bekе** +kambing.

Figure 17. Plus Symbol in Bugis-Indonesia Dictionary



bhondi (?)

boñjing sj alat musik atau bunyi-bunyian;
aboñjing memainkan *boñjing;*
paboñjing (pabunjing) tukang *boñjing*

yang bersangkutan) bernama Sutu

bsi besi

btĕng (?): *pada makapatih – ña* semuanya sebagai patih (pemimpin) **btĕng**-nya

buat *(bwat; wuat; wwat)* 1. bawa; bekerja; 2. karya, perbuatan;

Figure 18. Question Symbol in Bali Kuno-Indonesia Dictionary

3.3.6 The Template

As previously mentioned, we use a data template to make it available to process by the application. Our data template is specially formulated in excel. Before the data is converted to SQL, the data input staff must input the data through this format. The data input needs to follow the rules, or the template will get an error. The error template will not be able to upload to the application. As the data will be integrated into the Online KBBI, the template refers to the data structure on KBBI. This is the challenge to make various microstructure data from headword, pronunciation, usage label including abbreviation, lexicographic sign meets the standard of KBBI data structure. Some incomplete lexicographic information, such as the inexistence of part of speech and inexisted or incomplete definition, has also been a challenge we try to work for.

3.4 Handbook for Compiling the Bilingual Dictionary

The Language Agency already published a handbook for compiling the bilingual dictionary in 1990. This handbook provided information about data carding, data selecting, data presentation, orthography symbols, and a list of entry arrangements. Ideally, when compiling a dictionary, our lexicographers should refer to this book so that their dictionary has a similar format. Instead of it, the local language dictionary used for this research shows differences in a dictionary style, as we explained in section 3.3. Many factors may prevent them from using this book as a reference standard, such as lack of publication, out-of-date content, and having their style. Nevertheless, the differences will distress the lexicographer when inputting data because they have to equate the format.

4 Conclusion and Future Works

The digitalization of the dictionary is a significant project expected to make it easier for users to find word equivalents in regional languages. Not only that, this dictionary compilation application is expected to become a data bank for regional languages in Indonesia. This digitization project is inseparable from obstacles, as we have mentioned in section 3. These obstacles can be a reference for improvement for digitalization work that is currently still being carried out. We think about the opportunity to involve more communities or government and traditional institutions to know whether they have dictionaries published by the Language Agency, especially the old edition, to complete our dictionary collection.

Concerning the limited human resources, we can optimize the role of the dictionary staff, especially those in Kantor and Balai Bahasa, with training that can support dictionary digitalization. It can minimize the involvement of outsourcing or consultant outside the Agency. Even though this project has been completed, we still have competent staff in digitalization and data input verification.

Regarding the diversity of dictionary content, as we mentioned in section 3.3, the handbook of compiling bilingual dictionaries is currently being rewritten by adding a variety of up-to-date information. Findings related to differences in the presentation style of the dictionary microstructure when inputting data can be a suggestion for improvements or additions to the content of that book.

Moreover, we also propose two kinds of the technical instruction manual, namely manual of digitization that is more comprehensive and regularly updated for every people involved in this project that explain the details of application structure, lexicographical workflow, frequently found digitization cases that explain findings related to differences in the presentation style of the dictionary microstructure when inputting data. The second one is a manual for dictionary compilation works in Language Agency in the future. As every dictionary will be digitalized, this manual should provide information about microstructure style, data compilation technique, frequently found a case on the field and its solution. We hope that this idea will be able to uniform and urge consistency of presenting a local language dictionary that the Language Agency publishes in the future.

Last but not least is there is no perfect application. There will always be room for improvement, including the template. Some lexicographical components may not be accommodated by the template, such as scientific name, supporting picture, or illustration. It is a good improvement that the template can be updated to support those lexicographical components. A special column or special template is also needed to accommodate some words that are not available on KBBI. A special template to record words that need further research is also good to consider, especially those whose definition is inexisted or incomplete. The word can be documented on archived subsection to make the language researcher work easier. This whole idea may seem dreamy, but as Brian Tracy said, there are no limits on what we can achieve with our life, except the limits we accept in our minds.

5 References

Atkins, B. & Rundell, M. (2008). *The oxford guide to practical lexicography*. Oxford: Oxford University Press.

Hughes, L. M. (2004). *Digital collections: Strategic issues for the information manager*. London: Facet Publishing.

Janssen, M., Jansen, F., & Verkuyl, H. (2003). The Codification of Usage by Labels. In Sterkenburg, van (Eds.), *A practical guide to lexicography*. Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/tlrp.6.33ver

Klapicová, E. H. (2005). Composition of the entry in a bilingual dictionary. SKASE *Journal of Theoretical Linguistics*, 2(3), 57—74. http://www.skase.sk/Volumes/JTL04/05.pdf

Lexico. (n.d.). Usage label. In *Lexico.com dictionary*. Retrieved 25 Mei 2021 from https://www.lexico.com/definition/usage_label

Newmark, P. (1988). *A textbook of translation*. New York: Prentice Hall.

Rehg, K. L. (2018). Compiling dictionaries of endangered languages. In K. L. Rehg & L. Campbell (Eds.), *The handbook of endangered languages* (pp. 1—24). Oxford Handbook Online. DOI: 10.1093/oxfordhb/9780190610029.013.16

Wojowasito, T. (2007). Dasar-dasar leksikografi dwibahasa. In Dasar-Dasar Leksikologi dan Leksikografi [unpublished book]. Pusat Leksikografi dan Leksikolgi, Fakultas Ilmu Pengetahuan Budaya, Universitas Indonesia.

MORPHOLOGICAL INFORMATION OF LOANWORDS IN AN ETYMOLOGICAL DICTIONARY

Zahroh Nuriah, Totok Suhardijanto, Riska Risdiani
Universitas Indonesia
zahroh.nuriah@ui.ac.id; totok.suhardijanto@ui.ac.id; riskarisdiani@ui.ac.id

Abstract

An etymological dictionary is a source of information on the origin of words. As time goes by, the meaning of words can change and the morphological structures do so. Related to the phenomenon, a question arises: Is it necessary to put morphological information of loanwords in an etymological dictionary. In this research, the question is tried to be answered by analysing the loanwords of Dutch in Indonesian. Dutch is a language with rich affixes strictly bounded to part of speech, while Indonesian treats the part of speech of words more freely. Besides, the phonotactics of Dutch is more complex than Indonesian, resulting simplification of the syllable structure of Dutch loanwords in Indonesian. This paper focuses on the suffixes spelled in Indonesian as <asi>, <ir>, and <is>. Words with those suffixes are extracted from 2.000 loanwords from Dutch in Indonesian collected from Russel Jones (2008). The data is classified in line with the Dutch affixes. The meaning of the words is then checked in the Kamus Besar Bahasa Indonesia (KBBI) and the real use of those words is checked using the Indonesian corpus in *Sketch Engine*, namely the Indonesian WAC that consists of 109,236,814 tokens. By comparing the meaning of the loanwords in Dutch, in KBBI, and the meaning in the Indonesian context used in *Sketch Engine*, changes of meaning/use and differences are identified. The result shows that morphological information of loanwords is critically important to determine that two forms might have different meanings/uses. It can also help to understand the changes of loan affixes' distribution in the language.

Keywords: etymology, dictionary, morphological information, loanwords

1. Introduction

In the 19th century, the Dutch East Indies, now Indonesia, developed into a profitable colonial empire (Rijksmuseum, 2020). As the former colonist and colony, the Netherlands and Indonesia passed the long historical relationship from their perspectives through the generation by word of mouth, textbooks, institutions, traditions, monuments, and other documentation media. Moreover, both sides have a long relationship that resulted in shared pieces of knowledge and experiences in which loanwords took part. These loanwords revealed the etymological stories between two different cultures.

Language is changing across time, triggered by internal or external motivations. The history of language and words can be traced back through the study of etymology. Etymology is the investigation of word history (Durkin, 2009) and the result of the investigation is stored in an etymological dictionary. Each word has its own journey. However, a word cannot be secluded from its relationship with other words as a network and its constituents' morphemes if the word is polymorphemic. The core morpheme of a word, a free morpheme, or other morphemes, which is sometimes a bound morpheme, can develop on its own path.

With regard to Indonesian etymological dictionaries, it should be mentioned here some of the best works in Indonesian loanwords, such as Jones et al (2007) and De Casparis (1997). Not until Kamus Besar Bahasa Indonesia Fifth Edition (KBBI V) provided us with etymological information, it quite hard to find a dictionary with etymological information.

Unfortunately, KBBI V only provide etymological information in its web-based electronic format, and it is just available for registered users.

Words are adopted in their entirety, but when loanwords with certain affixes are borrowed a lot, there will be a parsing process that forms the word-formation rules in the donor or receiver language. Dutch is a language with rich affixes strictly bounded to part of speech, while Indonesian treats the part of speech of words more freely. Therefore, it is interesting to make Dutch loanwords in Indonesian as data in seeing the need for morphological information in an etymological dictionary. However, an interesting question comes later, such as is it necessary to put morphological information of loanwords in an etymological dictionary? This paper tries to answer the question by exploring and analysing Dutch loanwords in Indonesia with suffixes spelt in Indonesian as <asi>, <ir> and <is>.

2. Research Method

This paper examines the morpheme in the Dutch language loanwords, which is spelt as <asi>, <ir> and <is> in Indonesian. Words with those suffixes are extracted from 2,000 loanwords from Dutch in Indonesian collected from Russel Jones (2008). The data is classified in line with the Dutch affixes. Actually, the affixes have allomorphs, but in this paper we will focus on one form that has a similar form with other suffixes. If loanwords end with the same sound, but in Dutch are monomorphic words, these words are excluded from the data. The classification of the suffixes is as follows:

<asi>

Suffix	**Pronunciation**	**Meaning**	**Examples**
-atie	[asi]	noun	*amputasi, eliminasi*
-age	[aʒə]	noun	*bagasi, garasi*

Table 2.1

<ir>

Suffix	**Pronunciation**	**Meaning**	**Examples**
-eer	[er]	verb	*amputir, eliminir*
-uur	[yr]	object	*bordir, glasir*
-ier	[ir]	person	*amatir, bankir*

Table 2.2

<is>

Suffix	**Pronunciation**	**Meaning**	**Examples**
-isch	[ɪs]	adjective	*botanis*
-ist	[ɪst]	person	*anarkis, egois*
-je + s	[jəs]/ [is] + -s	diminutive plural	*ercis, hasyis*

Table 2.3

The Dutch wordform are checked following information of Jones (2008) and then the form and the meaning as well checked in *Van Dale Groot Woordenboek van de Nederlandse Taal* or abbreviated as VDGW (Boon, 2005). The meaning of the words in Indonesian is checked in *Kamus Besar Bahasa Indonesia* (KBBI), and the actual use of the words in Indonesian is then checked using the concordance of the Indonesian corpus in *Sketch Engine*, namely the Indonesian web corpus (IndonesianWaC) that consists of 109,236,814 tokens. By comparing the meaning of the loanwords in Dutch and the meaning in the Indonesian in KBBI and the context used in *Sketch Engine*, changes of form and meaning/use are identified.

3. Result

Based on the results of the analysis, suffixes <asi>, <ir> and <is> are homophonous morphemes that originated from several suffixes in Dutch. This occurs due to differences in the phonological system of Indonesian and Dutch, particularly the simplification of the phonotactic structure in Indonesian loanwords. Therefore, some Dutch sounds that do not exist in Indonesian are associated with the closest sound. The suffixes *-eer* [er], *-uur* [yr], *-eur* [ϕr], and *-ier* [ir] become [ir], since Indonesian does not have the vocal [y] and [ϕ]. Likewise, the sound suffix -*isch* [ɪs] and -*ist* [ɪst] both become -*is* [is], just like the suffix *-je+s* [is], because Indonesian does not recognise the consonant cluster 'st' as a coda. Moreover, Indonesian also does not distinguish the sound of a long vowel from a short vowel. Furthermore, the homophonous is caused naturally by the different sound association of some loanwords since one suffix is sometimes pronounced or spelt in Indonesian differently than what is suggested by Badan Pengembangan dan Pembinaan Bahasa. For example, the Dutch suffix *-eur* in *directeur* become [ur] *direktur* as suggested by Pusat Bahasa (2005)*,* but in *amateur* become [yr] *amatir.* Upon analysing the Russel Jones (2008) corpus, fifteen noticeable loanwords emerged, as shown in the table below.

Num.	Indonesian	Dutch	Van Dale	KBBI	Sketch Engine
1.	*amatir/ amatur*	*amateur*	person	noun/person	person / inanimate noun
					adjective
2.	*anarkis*	*anarchist*	person	person	person
					adjective
3.	*botanis*	*botanisch*	adjective	adjective	adjective
		botanist	person		person
4.	*eliminasi*	*eliminatie*	action noun	noun	abstract noun
5.	*eliminir*	*elimineer*	verb	> *eliminasi*	verb (pre-category)
6.	*eksploitasi*	*exploitatie*	noun	noun	abstract noun
7.	*eksploitir*	*exploiteer*	verb	> *eksploitasi*	verb (pre-category)
					person
8.	*grosir*	*grossier*	person	person	person
					abstract noun
9.	*egois*	*egoist*	person	person	person
					adjective
10.	*etnografis*	*etnografisch*	adjective	adjective	adjective
					person
11.	*filantropis*	*filantropisch*	adjective	adjective	adjective
					person
12.	*futuris*	*futurist*	person	person	person
					adjective
13.	*humanis*	*humanist*	person	person	person
					adjective
14.	*humoris*	*humorist*	person	person	person
					adjective
15.	*organis*	*organisch*	adjective	adjective	adjective
		organist	person	person	person

Table 3.1

From the observations on the concordance of the IndonesianWaC in Sketch Engine, there are some exciting data to look at more closely. Even though noun *amatir* is officially listed in KBBI, the form *amatur* is also shown in IndonesianWaC as a noun, and it is counted for 27 tokens. The homonym form of the <asi> suffix does not raise many issues among the original forms because all of them are morphemes that form inanimate nouns. For instance, as shown in Table 3.1, the Indonesian nouns *eliminasi* and *eksploitasi* were originated from Dutch action nouns *eliminatie* and noun *exploitatie*. However, the suffix <ir> is quite problematic because this suffix comes not only as a suffix that forms verbs but also nouns, both animate and inanimate. In the IndonesianWaC, the lemma *exploiteer* is used as a verb derived with *meng-* or *di-* as in the sentence (1) and as an animate noun in (2). This noun is not originated from Dutch, since the form for the person who does exploitation in Dutch is *exploitant.* Another example of the change of meaning/use can be seen in the Dutch lemma *grossier in* (3 and 4) that evolved as an inanimate noun. In sentence (3), *grosir* has no longer the meaning of a person 'wholesaler', but as the place to sell the commodities in quantity usually for resale, not for their own use and in sentence (4) as a way of wholesale trade, in a wholesale manner.

(1) *Akibatnya, rakyat pemilih sebagai konstituen, pemegang kedaulatan atas pilihan politik, diposisikan semata untuk di<u>eksploitir</u> sebagai alat legitimasi atas posisi kedudukan yang diraih elitnya.*

'As a result, the voters as constituents, holders of sovereignty over political choices, are positioned solely to be <u>exploited</u> as a means of legitimating the position of the position achieved by their elites.'

(2) *Perlu diatur secara tegas rambu-rambu HGU dalam UU karena sulitnya melakukan tindakan tegas bagi <u>eksploitir</u> air.*

'HGU signs need to be strictly regulated in the Act due to the difficulty of taking firmaction against water <u>exploiter</u>.'

(3) *Mereka umumnya pemilik <u>grosir</u> yang ingin mendapatkan harga lebih murah, memotong jalur distribusi dari para agen, yang berasal dari Jawa, Bandung, Sumatera, dan sebagainya.*

'They are commonly <u>wholesale</u> owners who desire to get lower prices and cut off distribution channels from agents who come from Java, Bandung, Sumatra, etc.'

(4) *Kawasan ini terkenal sebagai pusat perdagangan <u>grosir</u>, yang kemudian dikenal sebagai CBD (central business district) I Kota Surabaya.*

'This area is known as the centre of <u>wholesale</u> trade, which later popular as the CBD (central business district) I of Surabaya City.'

A further issue arose from some words with the suffix *-ir*: the suffix *-ir* also competes with the suffix *-asi*. As mentioned before, the Indonesian words with the suffix *-asi* like *eliminasi* and *eksploitasi* are nouns, since we can combine those words with the negation *bukan* that can only be attached to a noun, while *eliminir* and *eksploitir* are not nouns.

(5) *Ini bukan <u>eliminasi</u>.*

'This is not an <u>elimination</u>.'

(6) *Ini bukan <u>eksploitasi</u>.*

'This is not an <u>exploitation</u>.'

(7) **Ini bukan* <u>*eliminir*</u>*.*

'This is not <u>eliminate</u>.'

(8) **Ini bukan* <u>*eksploitir*</u>*.*

'This is not <u>exploit</u>.'

However, Indonesian words *eliminasi* and *eksploitasi* can have function as verbs, just like words with the suffix *-ir* borrowed from Dutch verbs like *elimineer* and *exploiteer*. Words with both suffixes, *-asi* and *-ir*, are treated the same since they are used as verbs in combination with the prefix *meng-* or *di-*.

The suffix <is> is also problematic because this suffix is a homophone since this form is originated from the Dutch suffix *-isch* that forms adjectives and from *-ist* that forms nouns with the meaning 'person' (9a & b). However, loanwords from words with the suffix *-ist* are founded as Indonesian adjectives (10a & b) as well.

(9a) *Tolstoy juga dikenal sebagai seorang* <u>*anarkis*</u>*.*

'Tolstoy was also known as an <u>anarchist</u>.'

(9b) *Mereka terkadang nampak bergerombol, tetapi bukan menyatu, karena pada dasarnya mereka hanya sekumpulan* <u>*egois*</u>*, yang kalau perlu bisa menginjak yang lain.*

'They sometimes appear to be in a bubble but not united because they are just a bunch of <u>egoists</u> who, if necessary, can look down on others.'

(10a) *Itu sudah kacau dan* <u>*anarkis*</u>*.*

'It's messed up and <u>anarchistic</u>.'

(10b) *Dia juga tidak boleh menjadi sosok yang individualis dan* <u>*egois*</u>*, hanya mementingkan diri sendiri, minim empati dan tidak mau melihat persoalan orang lain.*

'He is also not allowed to be individualistic and <u>egoistic</u>, put only himself as a priority, lack empathy, and does not want to see other people's problems.'

More interestingly, the Indonesian homonym *organis* refers to Dutch adjective *organisch* and Dutch noun *organist.* It seems that both words has *orgaan* as the stem, but actually, these words are loanwords. The adjective *organisch* is a loanword from German means having relation with an organ or a body part, while noun *organist* is from French *organiste* means the player of a musical instrument organ which is in Dutch named *orgel*. However, *orgaan* and *orgel* have the exact origin from the Latin *organum* means tool or instrument. Indonesian has borrowed both words with the same meaning in Dutch.

The use of the suffix <is> is also related to the English suffix *-ic*. In *Pedoman Umum Ejaan Bahasa Indonesia* (Tim Pengembang Pedoman Bahasa Indonesia, 2016:73), the English suffix *-ic* and the Dutch suffix *-isch* are related by suggesting to spell both affixes as *-ik*, while the English suffix *-ical* and again *-isc*h to be spelt as *-is*, while in *Pedoman Umum Pembentukan Istilah* (Pusat Bahasa, 2005: 18) is suggested to spell all suffixes, the Dutch suffix *-isch*, and English suffixes *-ic* and *-ical*, as <is>. The English suffix *-ic* that is parallel to the Dutch suffix *-iek* also poses another problem, since the suffix *-ic and -iek* can form nouns and adjectives. Nevertheless, in this paper we will not go deeper into the problem.

4. Discussion

The evolution of these morphemes involves changes in form related to phonological, morphological, syntactic, and semantic aspects. Loanwords are a result of a language contact that brings two language systems together. Loanwords from donor language with a complex phonotactic system by a receiver language with a simple phonotactic system will undoubtedly lead to simplifying the syllable structure. That is what happened to the Dutch loanwords in Indonesian. Suffixes in the loanwords with the spelling <asi>, <ir> and <is> are a simplification of various morphemes with complex syllable structures, as seen in the result section.

This phonotactics simplification then led to a change in meaning. The morpheme <is> is considered an adjective marker since the total loanwords with the suffix *-isch* is higher. The loanwords from Dutch words with the suffix *-ist* with the meaning of 'person' is not more than 30% of the words with the suffix *-isch*. Moreover, the morpheme <is> is considered an adjective marker because of the support of the English's hegemony after the Dutch language contact was reduced, since *-is* is considered borrowed from the English suffix *-ic*, which marks adjectives. However, in contrast to the *-isch* and *-ist* relationship in Dutch, the suffix *-ic* is pronounced differently from the suffix *-ist* in English, which also means 'person'. Thus, the meaning of <-is> as an adjective-forming suffix is stronger.

Apart from the phonotactic system, this morphological change also occurs due to the reinterpretation of the loanwords structure in the Indonesian language system itself. Due to the existence of homonym morphemes, such as <is> which has meaning as 'adjective' and 'person', some words are reinterpreted as to have both meanings. In the corpus, the word *egois* means not only 'person' but also as an adjective. The word *egois* as person cannot be categorized as a loanword from Dutch, nor English, because in Dutch, the adjective form is *egoistisch*, and in English, it is *egoistic*. It is, however, different to the case of the word *botanis*, which is also interpreted as 'person' and adjective. Although the problem is the same at first glance, the fact is that in Dutch, there are *botanisch* and *botanist* forms, both of which can be borrowed. As a result of the simplification of phonotactics, the form of the two words is the same, *botanis*.

Borrowing is not only in the form of loanwords but also affixes (Nuriah, 2013). The borrowing of words and affixes is not a separate process. However, the combination of words of a receiver language with foreign affixes is a shred of valid evidence that affixes are also borrowed. After being borrowed, foreign affixes undergo various changes due to differences in the structure of the receiver language and the donor language. The relationship between words and affixes is a network that is interconnected. The network is not only a relationship between a receiver language and a donor language, but also with another donor language. In this case, there is a network between Indonesian as a receiver language and Dutch and English as well as donor languages. And the loanwords of the donor language are possibly also loanwords from another languages. Moreover, there is also a discrepancy between the use of an affix by speaker and the plan of the formal institution for language planning, in this case Badan Pengembangan dan Pembinaan Bahasa. Based on this situation, morphological information for each loanword in the etymological dictionary is necessary.

Morphological information of loanwords is critically important to determine whether a form originated from one or more forms that have different meanings/uses. It can also help people to understand the form and meanings/uses changes of loan affixes in the language. This morphological information can also give insight into how the morphemes, the stems and the affixes as well, connect to each other in the network.

5. References

Boon, Ton den dan Dirk Geeraerts. (2005). *Van Dale Groot Woordenboek van de Nederlandse Taal.* Edisi XIV. Digitale Versie. Utrecht/Antwerpen: Van Dale Lexicografie B.V.

De Casparis, G. (1997). Sanskrit Loanwords in Indonesian: An Annotated Checklist of Words from Sanskrit in Indonesian and Traditional Malay. *NUSA Linguistic Studies of Indonesian and Other Languages in Indonesia 41*. Jakarta: Badan Penyelenggara Seri NUSA.

Durkin, P. (2009). *The Oxford Guide to Etymology*. OUP Oxford.

Jones, Russell, Carstairs Douglas, and Thomas Barclay. (2007) *Loan-words in Indonesian and Malay*. Leiden: KITLV Press.

Jones, R. (2008). *Loan-words in Indonesian and Malay*. Jakarta: KITLV–Yayasan Obor Indonesia.

Kamus Besar Bahasa Indonesia. Versi luring 2010-2011. http://ebsoft.web.id yang mengacu pada KBBI Daring (Edisi III) dan Versi daring http://pusatbahasa.kemdiknas.go.id/kbbi/].

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). *The Sketch Engine: Ten years on. Lexicography*, 1, 7–36. http://www.sketchengine.co.uk

Nuriah, Zahroh. (2013). "Loanwords Below Zero: Morfem Pinjaman *-(is)asi* dalam Bahasa Indonesia", dalam *Prosiding Seminar nasional Etimologi: "Teori dan Perkembangan Etimologi dalam Pelbagai Bahasa*, hlm. 133-140. Depok: Laboratorium Leksikologi dan Leksikografi Departemen Linguistik Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia.

Pusat Bahasa. (2005). *Pedoman Umum Pembentukan Istilah*. Jakarta: Pusat Bahasa Departemen Pendidikan Nasional.

Rijksmuseum. (2020, September 19). Retrieved from https://www.rijksmuseum.nl/en/rijksstudio/timeline-dutch-history/1820-1950- indonesia-and-decolonisation.

Tim Pengembang Pedoman Bahasa Indonesia. (2016). *Pedoman Umum Ejaan Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.

PENDEFINISIAN LEMA BIDANG ETNOMEDISIN DI LOMBOK BERDASARKAN KOMPONEN MAKNA

Rizki Gayatri, Menik Lestari
Regional Office for Language in West Nusa Tenggara Barat Province, Indonesia; Universitas Indonesia
rizki.gayatri@kemdikbud.go.id; menik.lestari01@ui.ac.id

Abstrak

Masyarakat Lombok, khususnya Suku Sasak, masih menjadikan tradisi pengobatan dengan tanaman obat sebagai cara untuk memperoleh hidup sehat (Museum Negeri Provinsi NTB, 2006: 1). Akan tetapi, dokumentasi mengenai leksikon etnomedisin ini sangat terbatas pada definisinya. Keterbatasan definisi tersebut dapat dilihat pada penelitian Museum Negeri Provinsi NTB (2006) dan Yamin, Dkk (2018) yang hanya membuat tabel dengan isian nama penyakit, bahan obat-obatan, cara pembuatan, dan cara pengobatan. Padahal, dokumentasi definisi yang komprehensif mengenai lema tanaman obat tersebut sangat diperlukan sebagai upaya pelestarian tradisi pengobatan masyarakat Sasak. Rumusan masalah penelitian ini adalah bagaimana wujud dan definisi lema bidang etnomedisin berdasarkan komponen maknanya melalui studi leksikografi dan semantik. Penelitian ini bertujuan untuk mendaftar dan mendefinisikan lema bidang etnomedisin di Lombok melalui kajian leksikografi dan semantik dalam bentuk kamus yang terbaca bagi masyarakat umum. Metode kualitatif dengan teknik studi pustaka dan wawancara menjadi metode penelitian dalam tulisan ini. Data diperoleh dari hasil identifikasi jenis tanaman obat tradisional dari Naskah *Takepan Usada* dan wawancara dengan masyarakat Sasak di Lombok. Kemudian, penelusuran lebih lanjut terkait gambaran fisik lema dilakukan melalui studi pustaka. Hasil penelitian menunjukkan bahwa ada 102 lema tanaman obat yang digunakan masyarakat Lombok. Adapun pendefinisian yang dilakukan diperoleh dari instrumen penelitian yang telah menerapkan analisis komponen makna Nida. Dalam instrumen, terdapat empat klasifikasi komponen makna, yakni bentuk tanaman, tinggi tanaman, bagian tanaman yang dijadikan obat, dan khasiat tanaman. Pengelompokkan lema-lema ini didasarkan atas fungsinya sebagai obat di sembilan lingkup pernyakit, yakni (1) penyakit di area kulit dan kelamin, (2) penyakit dalam, (3) penyakit tht, (4) penyakit di area gigi dan mulut, (5) penyakit di area mata, (6) penyakit saraf, (7) penyakit di area perut, (8) penyakit di area badan, dan (9) perawatan tubuh. Dalam proses pendefinisian, penyertaan nama latin tidak semua tercantum dalam lema karena tidak semua nama tanaman obat dalam bahasa Sasak ditemukan padanannya dalam bahasa Latin.

Kata Kunci: Definisi, lema, etnomedisin, tanaman obat, Sasak.

Pendahuluan

Masyarakat Lombok masih menjadikan tradisi pengobatan dengan tanaman obat sebagai cara untuk memperoleh hidup sehat (Museum Negeri Provinsi Nusa Tenggara Barat, 2006: 1). Hal ini dipengaruhi oleh ketersediaan tanaman obat tersebut. Pahrudin, dkk. (2020) menyatakan bahwa tumbuh-tumbuhan untuk pengobatan mudah di dapat di Pulau Lombok baik di sawah, kebun, maupun hutan. Meskipun demikian, pengobatan tradisional cukup bersaing dengan pesatnya kemajuan pengobatan modern. Bahan obat yang mudah didapatkan dan praktis menjadi daya tarik yang makin meminggirkan pengobatan tradisional di masyarakat Lombok.

Pahrudin, dkk. (2020) mengemukakan bahwa pengobatan tradisional Lombok dapat ditemukan secara lisan maupun tulisan. Sumber lisan dapat ditemukan di masyarakat dan sumber tulisan dapat ditemukan pada naskah-naskah kuno. Museum Negeri Provinsi Nusa Tenggara Barat pun pada tahun 2006 sebenarnya sudah berupaya melakukan dokumentasi nama-nama penyakit dan obat berdasarkan sumber dari Naskah *Takepan Usada*. Akan tetapi, hasil dokumentasi tersebut hanya berbentuk tabel yang berisi daftar nama penyakit dan tanaman obatnya. Padahal, penyajian dokumentasi yang detail dan terbaca bagi masyarakat umum menjadi hal yang sangat penting sebagai upaya pelestarian tradisi pengobatan tradisional masyarakat Lombok. Oleh sebab itu, diperlukan pendokumentasian lema tanaman obat yang komprehensif dan terbaca sebagai bentuk pelestarian bidang etnomedisin Lombok dari pesatnya perkembangan pengobatan modern.

Penelitian ini merupakan penelitian lanjutan dari studi etnomedisin di Lombok. Sebelumnya, Museum Negeri Provinsi NTB pada tahun 2006 telah membuat dokumentasi nama-nama penyakit dan obat berdasarkan sumber dari *Takepan Usada*. Lontar kuno yang kini menjadi koleksi Museum Negeri Provinsi Nusa Tenggara Barat ini menghimpun tradisi pengobatan tradisional dalam wujud deskripsi naskah dan tabel dengan isian nama penyakit, bahan obat- obatan, cara pembuatan, dan cara pengobatan. Kedua, ada penelitian dari Yamin, dkk (2018) berjudul *Pengobatan dan Obat Tradisional Suku Sasak di Lombok* yang mengambil data dari penelitian pertama oleh Museum Negeri Provinsi Nusa Tenggara Barat. Penelitian tersebut menjelaskan daftar nama tanaman obat dengan metode penelitian linguistik-antropologi. Hampir sama dengan penelitian dari museum, perbedaannya dalam tulisan ini, Yamin, dkk (2018) menyertakan kutipan transliterasi naskah untuk beberapa tanaman obat dan menambah daftar nama tanaman obat dengan nama tanaman yang diperoleh dari wawancara di masyarakat. Selanjutnya ada penelitian dari Arrozi, dkk (2020) membahas *Leksikon Etnomedisin dalam Pengobatan Tradisional Sasak*: *Kajian Antropolinguistik*. Dalam penelitian tersebut, Arrozi, dkk (2020) memaparkan karakteristik beberapa leksikon tanaman obat yang diteliti beserta nama penyakitnya tanpa menyertakan definisi detail.

Penelitian-penelitian sebelumnya telah mendokumentasikan tanaman obat ke dalam bentuk karya ilmiah. Akan tetapi, hampir semua tulisan tidak membuat dokumentasi dalam wujud definisi utuh. Dengan kata lain, bentuk dokumentasi bahasa yang mudah dipahami masyarakat belum tersedia. Oleh sebab itu, penelitian ini berusaha membuat wujud dokumentasi nama tanaman obat khususnya di Pulau Lombok dengan menggunakan kajian leksikografi dan semantik.

Penelitian ini berfokus pada masalah dokumentasi lema dan definisi tanaman obat masyarakat Lombok. Adapun pertanyaan penelitian ini, yaitu bagaimana wujud dan definisi lema bidang etnomedisin berdasarkan komponen maknanya melalui studi leksikografi dan semantik. Ancangan leksikografi dan semantik dipilih dalam penelitian ini untuk memberikan definisi yang komprehensif bagi masyarakat umum. Dari pertanyaan penelitian tersebut, tujuan penelitian ini adalah memberikan bentuk dokumentasi lema dan definisi bidang etnomedisin di Lombok dalam bentuk kamus yang terbaca bagi masyarakat umum.

Metode Penelitian

Metode kualitatif dengan teknik studi pustaka dan wawancara menjadi metode penelitian dalam tulisan ini. Metode kualitatif dipilih karena penelitian ini berfokus untuk membedah secara mendalam definisi lema bidang etnomedisin di Lombok melalui analisis komponen makna Nida (1975) dan mendokumentasikannya menggunakan ancangan leksikografi dalam bentuk kamus yang terbaca bagi masyarakat umum. Sumber data penelitian ini adalah tabel identifikasi jenis tanaman obat tradisional dari Naskah *Takepan Usada* dari Museum Negeri Provinsi Nusa Tenggara Barat. Selain itu, terdapat juga data pendukung berupa hasil wawancara dengan masyarakat Sasak di Lombok mengenai tanaman obat. Data pendukung tersebut berkaitan dengan studi etnomedisin yang mengungkapkan pengetahuan lokal berbagai etnis dalam menjaga kesehatannya. Selain itu, penelitian ini juga menggunakan metode studi pustaka untuk mengumpulkan data pendukung berupa gambaran fisik dari tanaman.

Penelitian ini menggunakan ancangan interdisipliner, yakni etnomedisin, leksikografi, dan semantik dalam pengolahan data. Adapun langkah kerja untuk menganalisis data adalah sebagai berikut.

1. Studi pustaka dari Naskah *Takepan Usada*. Museum Negeri Provinsi Nusa Tenggara Barat telah merangkum nama tanaman obat ke dalam buku yang berjudul *Obat-obatan Tradisional Lombok.* Dalam buku, museum telah menunjukkan jenis- jenis penyakit yang dalam penelitian ini juga dijadikan dasar pengelompokkan lema dalam instrumen penelitian.
2. Melakukan wawancara kepada masyarakat yang ada di Lombok Timur dan Lombok Barat sebagai metode etnomedisin. Data wawancara juga disesuaikan dengan temuan dari *Pengobatan dan Obat Tradisional Suku Sasak di Lombok* yang disusun Yamin, dkk (2018). Selain itu, beberapa lema yang belum jelas komponen maknanya dilengkapi keterangannya berdasarkan sumber dari Kamus Besar Bahasa Indonesia Daring.
3. Instrumen penelitian yang berisi lema etnomedisin masyarakat sasak lombok disajikan sebagai bagian dari penelitian leksikografi. Dalam menyusun instrumen ini, lema akan dikelompokkan berdasarkan lingkup jenis penyakit. Dalam satu kelompok penyakit, akan ditampilkan lema dan komponen maknanya. Komponen makna dalam tabel tersebut berisi bentuk tanaman, tinggi tanaman, bagian tanaman yang dijadikan obat, dan khasiat tanaman. Penentuan komponen makna ini disesuaikan dengan karakteristik umum yang ada pada semua jenis tanaman obat dan komponen yang dapat dijadikan pembanding antara satu tanaman dan tanaman obat lainnya.
4. Kemudian, definisi dibuat berdasarkan komponen makna dari Nida (1975) yang telah dirangkum dalam instrumen penelitian. Kemudian, semua definisi lema dijadikan satu dan diurutkan berdasarkan abjad sesuai dengan ancangan leksikografi. Data tambahan untuk lema berupa nama Latin tanaman juga disertakan saat pendefinisian lema.

Hasil

Bagian ini memaparkan instrumen penelitian yang dijadikan alat untuk menyusun definisi tanaman obat. Secara umum, instrumen penelitian tersebut menggunakan analisis komponen makna Nida (1975) yang dikategorisasikan dengan lingkup penyakitnya sesuai dengan klasifikasi jenis penyakit yang ada pada Naskah *Takepan Usada.* Adapun kategorisasi tanaman obat tersebut dibedakan atas 9 jenis penyakit, yakni (1) penyakit di area kulit dan kelamin, (2) penyakit dalam, (3) penyakit tht, (4) penyakit di area gigi dan mulut, (5) penyakit di area mata, (6) penyakit saraf, (7) penyakit di area perut, (8) penyakit di area badan, dan (9) perawatan tubuh. Pemisahan kategori tersebut dilakukan untuk memudahkan proses analisis komponen makna. Selain itu, komponen makna tanaman obat yang dijadikan pembanding adalah bentuk tanaman, tinggi tanaman, bagian tanaman yang dijadikan obat, dan khasiat tanaman. Komponen bentuk tanaman dibagi atas perdu, pohon, tumbuhan, semak, dan terna. Bentuk tanaman berupa tumbuhan merangkum tumbuhan menjalar, tumbuhan merambat pada tumbuhan lain, dan tumbuhan air. Adapun kategori terna memuat jenis umbi dan tanaman herbal. Komponen bagian tanaman yang dijadikan obat dibedakan atas biji, daun, batang, bunga, akar, dan buah. Selain itu, khasiat tanaman diperoleh dari data pendukung studi pustaka dan wawancara masyarakat setempat. Meskipun proses analisis komponen makna ini dibedakan atas lingkup jenis penyakitnya, pendefinisian lema tanaman obat dibuat dengan mengikuti ancangan leksikografi, yakni definisi yang diurutkan berdasarkan abjad dan disertakan keterangan tambahan yang diperoleh saat wawancara. Adapun tabel hasil analisis komponen makna tanaman obat bidang etnomedisin Lombok adalah sebagai berikut.

1. Tanaman obat untuk penyakit di area kulit dan kelamin

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	arak	-	+	-	-	-	+	-	+	-	-	-	-	penyakit kulit
2.	bawang merah	-	-	-	-	+	+	+	-	-	-	-	+	eksem
3.	belimbing wuluh	-	+	-	-	-	-	+	+	-	-	-	-	bisul
4.	beru	-	+	-	-	-	-	-	-	-	-	-	-	bisul
5.	bila	-	+	-	-	-	-	-	+	-	-	-	-	eksem
6.	birak	-	-	+	-	-	+	-	-	-	+	-	-	biduran
7.	brotowali	-	+	-	-	-	-	-	-	+	-	-	-	gatal-gatal
8.	cemara	-	+	-	-	-	-	-	+	-	-	-	-	kulit bersisik
9.	delima	+	-	-	-	-	+	-	-	-	-	-	+	cacar
10.	gerepek	-	+	-	-	-	-	-	+	-	-	-	-	cacar
11.	kanangas	-	-	-	+	-	-	-	-	-	-	+	-	eksem
12.	kangkung	-	-	+	-	-	+	-	+	-	-	-	-	biduran
13.	kapas	-	-	-	+	-	+	-	-	-	-	-	+	luka bakar
14.	alang-alang	-	-	-	+	-	+	-	-	-	-	+	-	sifilis
15.	daun sendok	-	-	-	-	+	+	-	-	-	-	+	-	keputihan
16.	klokos udang	-	+	-	-	-	-	-	-	-	-	-	+	kulit bersisik
17.	ketimbus	-	+	-	-	-	-	-	-	+	-	-	-	sakit koreng raja
18.	laos	-	-	-	-	+	+	-	-	+	-	-	-	bisul dan panuan
19.	lembain	-	-	-	-	+	+	-	+	-	-	-	-	biduran dan eksem
20.	pakis	+	-	-	-	-	+	-	+	-	-	-	-	biduran
21.	pandan	-	+	-	-	-	-	-	+	-	-	-	-	penyakit kulit
22.	petai	-	+	-	-	-	-	-	-	-	-	-	+	borok
23.	suren	-	+	-	-	-	-	-	-	+	-	-	-	bisul
24.	tembakau	-	+	-	-	-	+	-	+	-	-	-	-	borok

2. obat untuk untuk penyakit dalam

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	bebele	-	-	+	-	-	+	-	+	-	-	-	-	kencing nanah
2.	dadap serep	-	+	-	-	-	-	-	+	-	-	-	-	kencing batu
3.	gandarusa	+	-	-	-	-	+	-	+	-	-	-	-	kencing batu
4.	bidara putih	-	+	-	-	-	-	-	-	-	-	+	-	cuci darah
5.	brotowali	-	-	+	-	-	+	-	-	+	-	-	-	kolesterol
6.	daun dewa	-	-	-	-	+	+	-	+	-	-	-	-	kolesterol
7.	kelapa	-	+	-	-	-	-	-	-	-	-	-	+	darah tinggi, ginjal

8.	semanggi	-	-	+	-	-	+	-	+	-	-	-	-	kencing manis
9.	semet meong	+	-	-	-	-	+	-	+	-	-	-	-	ginjal
10.	meniran	-	-	-	-	+	+	-	-	+	-	-	-	ginjal
11.	petikan kebo	-	-	-	-	+	+	-	+	-	-	-	-	ginjal
12.	re	-	-	-	+	-	+	-	-	-	-	+	-	kencing batu
13.	selederi	-	-	-	-	+	+	-	+	-	-	-	-	darah tinggi
14.	sambiloto	-	-	-	-	+	+	-	+	-	-	-	-	kencing manis
15.	songgo langit	-	-	-	+	-	+	-	+	-	-	-	-	ginjal
16.	tapak dara	+	-	-	-	-	+	-	+	-	-	-	-	kencing manis

3. obat untuk penyakit tht

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	bangle	-	-	-	-	+	+	-	-	+	-	-	-	radang
2.	bebele	-	-	+	-	-	+	-	+	-	-	-	-	gondok
3.	kecubung	+	-	-	-	-	+	+	-	-	-	-	-	panas dalam
4.	kemangi	-	-	-	-	+	+	-	+	-	-	-	-	sakit telinga

4. obat untuk penyakit di area gigi dan mulut

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	barora	-	+	-	-	-	-	-	+	-	-	-	-	sariawan
2.	kelapa	-	+	-	-	-	-	-	-	-	-	-	+	sariawan
3.	pandan	-	+	-	-	-	-	-	+	-	-	-	-	sariawan
4.	pecut kuda	-	-	-	-	+	+	-	+	-	-	-	-	amandel
5.	pinang	-	+	-	-	-	-	+	-	-	-	-	-	penghilang bau mulut
6.	sager	+	-	-	-	-	+	-	+	-	-	-	-	sariawan
7.	salam	-	+	-	-	-	-	-	+	-	-	-	-	sariawan
8.	sirih	-	-	+	-	-	+	-	+	-	-	-	-	penghilang bau mulut
9.	wareng	-	+	-	-	-	-	-	-	-	-	-	+	gusi berdarah

5. obat untuk penyakit di area mata

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	bageq	-	+	-	-	-	-	-	-	-	-	-	+	sakit mata
2.	bambu	-	+	-	-	-	-	-	-	+	-	-	-	sakit mata
3.	delima	+	-	-	-	-	+	-	-	-	-	-	+	sakit mata
4.	peria	-	-	+	-	-	+	-	-	-	-	-	+	sakit mata
5.	petikan kebo	-	-	-	-	+	+	-	+	-	-	-	-	sakit mata

6. obat untuk penyakit saraf

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	belimbing wuluh	-	+	-	-	-	-	-	+	-	-	-	-	rematik
2.	cabai hutan	-	-	+	-	-	+	-	-	-	-	-	+	pegal linu
3.	entut-entut	-	-	+	-	-	+	-	+	-	-	-	-	sakit pinggang
4.	kemuning	-	+	-	-	-	-	-	+	-	-	-	-	rematik
5.	sekuh	-	-	-	-	+	+	-	-	+	-	-	-	pegal linu
6.	terung	+	-	-	-	-	+	-	-	-	-	-	+	rematik

7. obat untuk untuk penyakit area perut

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	acar	-	+	-	-	-	-	+	-	-	-	-	-	desentri
2.	adas	-	-	-	-	+	+	+	-	-	-	-	-	desentri
3.	bunut	-	+	-	-	-	-	-	+	-	-	-	-	desentri
4.	cemara	-	+	-	-	-	-	-	+	-	-	-	-	kejang perut
5.	delima	+	-	-	-	-	+	-	-	-	-	-	+	berak darah
6.	entut-entut	-	-	+	-	-	+	-	+	-	-	-	-	sakit perut
7.	jambu biji	-	+	-	-	-	-	-	+	-	-	-	-	sakit perut
8.	jarak pagar	+	-	-	-	-	+	-	+	-	-	-	-	menceret
9.	jowet	-	+	-	-	-	-	-	-	-	-	-	+	menceret
10.	kacang hijau	+	-	-	-	-	+	-	-	-	+	-	-	muntaber
11.	kates	-	+	-	-	-	-	-	-	-	+	-	-	menceret
12.	kunyit	-	-	-	-	+	+	-	-	+	-	-	-	sakit perut
13.	lada	-	-	+	-	-	-	-	-	-	-	-	+	perut
14.	lita	-	+	-	-	-	-	-	-	+	-	-	-	mag
15.	merang	-	-	-	-	+	+	-	-	+	-	-	-	sakit perut

16.	peko	-	+	-	-	-	-	-	+	-	-	-	-	sakit perut
17.	raju mas	-	+	-	-	-	-	-	-	-	-	-	+	diare
18.	randu	-	+	-	-	-	+	-	+	-	-	-	-	diare
19.	rembiga	-	+	-	-	-	+	-	-	-	-	+	-	cacingan
20.	wortel	-	-	-	-	+	+	-	-	-	-	-	+	desentri

8. obat untuk untuk penyakit area badan

No.	Nama Tanaman Komponen makna	bentuk tanaman						bagian tanaman yang dijadikan obat						khasiat tanaman
		perdu	pohon	tumbuhan	semak	terna	tinggi kurang dari 5 m	bijinya	daunnya	batangnya	bunganya	akarnya	buahnya	
1.	acar	-	+	-	-	-	-	+	-	-	-	-	-	demam
2.	adas	-	-	-	-	+	+	+	-	-	-	-	-	demam
3.	aiq-aiq	-	+	-	-	+	+	-	+	-	-	-	-	demam
4.	alang-alang	-	-	-	+	-	+	-	-	-	-	+	-	demam
5.	arak	-	-	-	+	-	+	-	+	-	-	-	-	demam
6.	bengkel	-	+	-	-	-	-	-	-	+	-	-	-	sakit ngilu
7.	bidara putih	-	+	-	-	-	-	-	-	-	-	+	-	demam
8.	bokah	-	-	+	-	-	+	-	+	-	-	-	-	demam pada bayi
9.	bunga sepatu	+	-	-	-	-	+	-	-	-	+	-	-	panas
10.	bunut	-	+	-	-	-	-	-	+	-	-	-	-	kejang karena panas
11.	inggu	-	-	-	-	+	+	-	+	-	-	-	-	sakit kepala sebelah
12.	jarak pagar	+	-	-	-	-	-	-	+	-	-	-	-	mimisan, panas
13.	kenanga	-	+	-	-	-	-	-	-	-	+	-	-	panas
14.	kesambiq	-	+	-	-	-	-	-	+	-	-	-	-	pusing tujuh hari
15.	kesembung	+	-	-	-	-	+	-	+	-	-	-	-	lemas
16.	ketumbar	+	-	-	-	-	+	+	-	-	-	-	-	panas
17.	lada	-	-	+	-	-	-	-	-	-	-	-	+	malaria
18.	lempuyang	-	-	-	-	+	+	-	-	+	-	-	-	anemia
19.	paoq	-	+	-	-	-	-	-	+	-	-	-	-	silu
20.	sambung nyawa	-	-	-	-	+	+	-	+	-	-	-	-	pusing
21.	selasih hitam	-	-	-	-	+	+	+	-	-	-	-	-	panas
22.	sereto	+	-	-	-	-	-	-	-	+	-	-	-	malaria
23.	serikaya	+	-	-	-	-	-	-	-	-	-	-	+	malaria
24.	soka	-	+	-	-	-	+	-	-	+	-	+	-	panas
25.	temulawak	-	-	-	-	+	+	-	-	+	-	-	-	flu dan demam
26.	waru laut	-	+	-	-	-	-	-	+	-	-	-	-	silu

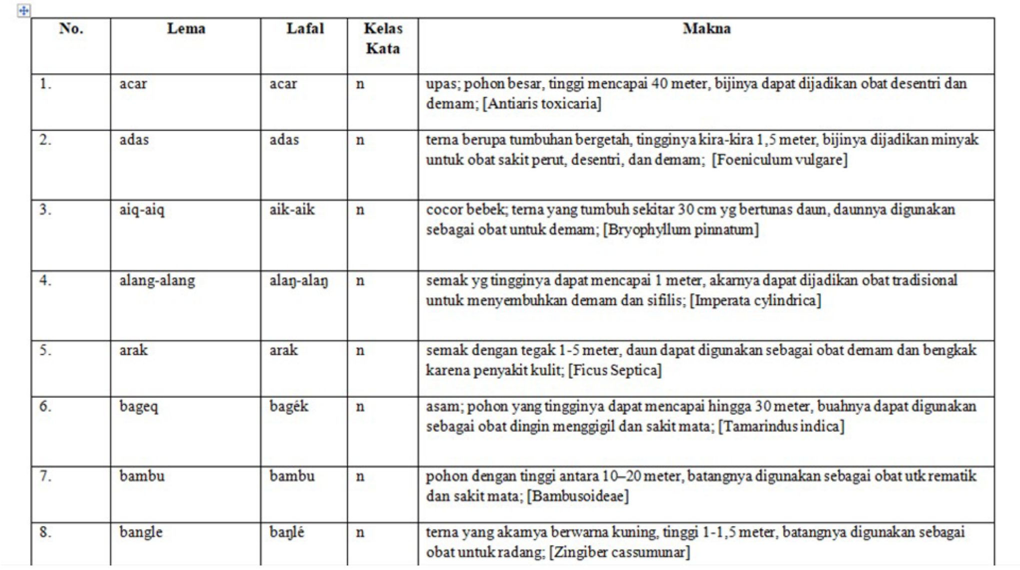
9. obat untuk perawatan tubuh														
No.	**Nama** **Tanaman** **Komponen makna**	**bentuk tanaman**						**bagian tanaman yang dijadikan obat**						**khasiat tanaman**
		perdu	**pohon**	**tumbuhan**	**semak**	**terna**	**tinggi kurang dari 5 m**	**bijinya**	**daunnya**	**batangnya**	**bunganya**	**akarnya**	**buahnya**	
1.	jeringo	-	-	-	-	+	+	-	-	-	-	+	-	penangkal gangguan makhlus halus
2.	keroton	+	-	-	-	-	+	-	-	-	+	-	-	untuk membuat rambut tebal
3.	kopi	-	+	-	-	-	+	+	-	-	-	-	-	menyegarkan kulit
4.	kuluh	-	+	-	-	-	-	-	-	-	-	-	+	menyehatkan badan
5.	kelor	-	+	-	-	-	-	-	+	-	-	-	-	menyehatkan badan
6.	lebui	+	-	-	-	-	+	-	-	-	-	-	+	tambah darah
7.	lekong	-	+	-	-	-	-	-	-	-	-	-	+	untuk membuat rambut hitam
8.	lobak	-	-	-	-	+	+	-	-	-	-	-	+	melancarkan persalinan
9.	nenas	-	-	-	-	+	+	-	-	-	-	-	+	melancarkan haid
10.	tebu besi	-	-	-	+	-	+	-	-	+	-	-	-	anti ubanan
11.	ketujur	-	+	-	-	-	-	-	+	-	-	-	-	melancarkan air susu

Analisis

Tabel analisis komponen menunjukkan bahwa terdapat 102 lema yang masuk dalam tanaman obat di Lombok berdasarkan lingkup penyakitnya. Secara umum, tanaman obat yang berkhasiat untuk penyakit area badan menempati lingkup dengan jumlah terbanyak dibandingkan dengan lingkup tanaman obat untuk penyakit lain, yakni 26 lema. Hal tersebut dimungkinkan karena penyakit tersebut memang paling sering dialami oleh masyarakat setempat. Berdasarkan wawancara dan data, penggunaan tanaman obat sebagai bagian dari studi etnomedisin memerlukan pengetahuan tinggi terkait cara pengolahan dan dosis penggunaan. Sementara itu, tabel komponen makna juga menunjukkan bahwa secara umum tanaman obat di Lombok berbentuk pohon dan tinggi tanaman paling banyak kurang dari 5 meter, serta bagian tanaman yang paling banyak dimanfaatkan sebagai obat adalah daun. Tanaman berbentuk pohon berjumlah 51 lema, tinggi tanaman yang kurang dari 5 meter berjumlah 70 lema, dan bagian tanaman yang dijadikan obat berupa daun berjumlah 53 lema.

Dalam komponen makna tumbuhan, kategori tersebut dapat berupa tanaman menjalar, tumbuhan yang menempel pada tumbuhan lain, dan tumbuhan air. Sementara itu, komponen makna terna dalam tabel komponen makna memuat kelompok umbi-umbian dan juga tanaman herbal. Komponen makna bentuk tanaman ternyata berkaitan dengan komponen makna lain, yakni tinggi tanaman. Untuk tanaman yang merupakan semak atau terna, data menunjukkan bahwa tinggi tanaman tersebut dapat dipastikan di bawah 5 meter. Sementara untuk tanaman berupa pohon kebanyakan memiliki tinggi di atas 5 meter. Ada beberapa tanaman yang masuk ke dalam kategori pohon, tetapi memiliki tinggi di bawah 5 meter. Selain itu, komponen makna bagian tanaman yang dijadikan obat menunjukkan bahwa bagian batang dan akar saling berkaitan. Artinya, apabila suatu penyakit dapat disembuhkan dengan rebusan akar, beberapa penyakit tersebut juga dapat disembuhkan dengan mengombinasikan rebusan akar dan batang.

Berdasarkan pemaparan di atas, tabel analisis komponen makna ini memang sangat penting untuk membuat definisi yang komprehensif dan konsisten pada lema-lema tanaman obat. Komponen-komponen makna yang ada pada table analisis komponen didokumentasikan melalui kajian leksikografi dalam bentuk kamus yang terbaca bagi masyarakat umum. Kamus tanaman obat Lombok ini menggunakan bahasa Indonesia dengan informasi pelafalan, kelas kata, dan definisi yang disertakan nama latinnya. Penulisannya pun disesuaikan dengan urutan abjad seperti pada kamus umumnya. Adapun contoh tampilan kamusnya adalah sebagai berikut.



No.	Lema	Lafal	Kelas Kata	Makna
1.	acar	acar	n	upas; pohon besar, tinggi mencapai 40 meter, bijinya dapat dijadikan obat desentri dan demam; [Antiaris toxicaria]
2.	adas	adas	n	terna berupa tumbuhan bergetah, tingginya kira-kira 1,5 meter, bijinya dijadikan minyak untuk obat sakit perut, desentri, dan demam; [Foeniculum vulgare]
3.	aiq-aiq	aik-aik	n	cocor bebek; terna yang tumbuh sekitar 30 cm yg bertunas daun, daunnya digunakan sebagai obat untuk demam; [Bryophyllum pinnatum]
4.	alang-alang	alaŋ-alaŋ	n	semak yg tingginya dapat mencapai 1 meter, akarnya dapat dijadikan obat tradisional untuk menyembuhkan demam dan sifilis; [Imperata cylindrica]
5.	arak	arak	n	semak dengan tegak 1-5 meter, daun dapat digunakan sebagai obat demam dan bengkak karena penyakit kulit; [Ficus Septica]
6.	bageq	bagék	n	asam; pohon yang tingginya dapat mencapai hingga 30 meter, buahnya dapat digunakan sebagai obat dingin menggigil dan sakit mata; [Tamarindus indica]
7.	bambu	bambu	n	pohon dengan tinggi antara 10–20 meter, batangnya digunakan sebagai obat utk rematik dan sakit mata; [Bambusoideae]
8.	bangle	baŋlé	n	terna yang akarnya berwarna kuning, tinggi 1-1,5 meter, batangnya digunakan sebagai obat untuk radang; [Zingiber cassumunar]

Simpulan

Masyarakat Lombok masih menggunakan tanaman obat sebagai pengobatan tradisional di daerahnya. Hal ini ditunjukkan dengan adanya temuan 102 tanaman obat yang digunakan oleh masyarakat setempat. Dari 102 lema tanaman obat tersebut, tanaman obat yang digunakan untuk mengobati penyakit area badan menjadi lingkup tanaman obat yang paling banyak, yakni 26 lema. Tingginya jumlah lema tersebut dimungkinkan karena penyakit area badan paling sering dialami oleh masyarakat setempat. Selain itu, apabila dilihat pada komponen bentuk tanaman, bentuk pohon menjadi bentuk yang paling banyak ditemukan pada tanaman obat Lombok dan tingginya ada yang kurang dari 5 meter. Selain itu, bagian tanaman yang dijadikan obat di Lombok paling banyak adalah daun tanaman obatnya. Data juga menunjukkan bahwa adanya keterkaitan antara bagian akar dan batang dari tanaman obat di Lombok.

Tabel analisis komponen makna yang dikemukakan Nida (1975) memberikan batasan yang jelas antara tanaman obat yang satu dengan tanaman obat yang lain. Tabel tersebut memperlihatkan dengan jelas perbandingan setiap komponen makna dari lema-lema tanaman obat yang ada pada satu lingkup penyakit yang sama. Hal tersebut tentunya memudahkan masyarakat dalam mengolah tanaman obat sehingga kesalahan pengolahan tanaman obat dapat diminimalisasi. Selain itu, tabel analisis komponen makna juga memudahkan pembuatan definisi lema tanaman obat yang konsisten dan komprehensif. Dengan demikian, dokumentasi pembuatan kamus tanaman obat di Lombok dengan menggunakan analisis komponen Nida (1975) dan ancangan leksikografi ini diharapkan dapat menyajikan kamus yang terbaca bagi masyarakat umum dan menjadi upaya pelestarian kebudayaan masyarakat Lombok.

Referensi

Aminuddin. (2008). *Semantik Pengantar Studi Tentang Makna.* Bandung: Sinar Baru Algesindo.

Arrozi, P., Burhanuddin., & Saharudin. (2020). Leksikon Etnomedisin dalam Pengobatan Tradisional Sasak: Kajian Antropolinguistik. *Jurnal Mabasan: Masyarakat Bahasa & Sastra Nusantara,* 14(1), 17—30. doi:10.26499/mab.v14i1.308.

Aridawati, I.A., Thoir, N., Purwa, I.M., dan Sutana, Dwi. (1995). *Struktur Bahasa Sasak Umum.* Jakarta: Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan Badan Pengembangan dan Pembinaan Bahasa. (2019). Bahasa di Provinsi Nusa Tenggara Barat. Provinsi Nusa Tenggara Barat - Peta Bahasa (kemdikbud.go.id)

Cruse, D.A. (1995). *Lexical Semantics.* New York: Cambridge University Press.

Jackson, H. (2013). *The Bloomsbury Companion to Lexicography.* New York, USA: Bloomsbury Publishing Plc.

Kantor Bahasa Nusa Tenggara Barat. (2016). *Ensiklopedia Bahasa Sasak.* Mataram: Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan. Kridalaksana, H. (1984). *Kamus Linguistik*. Jakarta: P.T. Gramedia.

Kridalaksana, H. (2010). *Sendi-sendi Ilmiah bagi Pembinaan Bahasa*. Jakarta: Laboratorium Leksikologi dan Leksikografi.

Mahsun. (1997, July 21-24). Kekerabatan Bahasa-Bahasa di NTB Kajian Tanah Asal Penutur- Penuturnya. Pro-Mahsun. Kekerabatan Bahasa-Bahasa di NTB - Prof. Dr. Mahsun, M.S. (prof- mahsun.com).

Museum Negeri Provinsi Nusa Tenggara Barat. (2006). *Obat-Obatan Tradisional Lombok.* Mataram: Pemerintah Provinsi Nusa Tenggara Barat, Dinas Kebudayaan dan Pariwisata, Museum Negeri Provinsi Nusa Tenggara Barat.

Nida, E.A. (1975). *Componential Analysis of Meaning*. Netherland: The Hague.

Pemerintah Daerah Provinsi Nusa Tenggara Barat. (2019). *Peraturan Gubernur Nusa Tenggara Barat Nomor 19 Tahun 2019 tentang Rencana Kerja Pemerintah Daerah Provinsi Nusa Tenggara Barat Tahun 2020.* Mataram: Pemerintah Daerah Provinsi Nusa Tenggara Barat.

Santoso, T. (2015). Komponen Makna Kata 'Mencuri/Mengambil' dalam Bahasa Indonesia. *Medan Makna* 13(1), 55—61.

Silalahi, M. (2016). Studi Etnomedisin di Indonesia dan Pendekatan Penelitiannya. *JDP* 9(3), 117—124.

Supriyanti, N. (2012). Praktik Leksikografi atas Nomina Persona Berorientasi Gender dalam Kamus Besar Bahasa Indonesia Edisi IV. Program Pascasarjana Program Studi Linguistik, Universitas Indonesia. Diakses dari https://drive.google.com/file/d/1Cp8GXjxwFAJQNBmY- soIPdqp52rISC4Z/view?usp=sharing.

Yamin, M., Burhanudin., Jamaluddin., & Nasruddin. (2018). Pengobatan dan Obat Tradisional Suku

Sasak di Lombok. *Jurnal Biologi Tropis,* 18(1), 1-12. doi:10.29303/JBT.V18I1.463.

Lampiran Daftar Lema dan Definisi Tanaman Obat Lombok

No.	Lema	Lafal	Kelas Kata	Makna
1.	acar	acar	n	upas; pohon besar, tinggi mencapai 40 meter, bijinya dapat dijadikan obat desentri dan demam; [Antiaris toxicaria]
2.	adas	adas	n	terna berupa tumbuhan bergetah, tingginya kira-kira 1,5 meter, bijinya dijadikan minyak untuk obat sakit perut, desentri, dan demam; [Foeniculum vulgare]
3.	aiq-aiq	aik-aik	n	cocor bebek; terna yang tumbuh sekitar 30 cm yg bertunas daun, daunnya digunakan sebagai obat untuk demam; [Bryophyllum pinnatum]

4.	alang-alang	alaŋ-alaŋ	n	semak yg tingginya dapat mencapai 1 meter, akarnya dapat dijadikan obat tradisional untuk menyembuhkan demam dan sifilis; [Imperata cylindrica]
5.	arak	arak	n	semak dengan tegak 1-5 meter, daun dapat digunakan sebagai obat demam dan bengkak karena penyakit kulit; [Ficus Septica]
6.	bageq	bagék	n	asam; pohon yang tingginya dapat mencapai hingga 30 meter, buahnya dapat digunakan sebagai obat dingin menggigil dan sakit mata; [Tamarindus indica]
7.	bambu	bambu	n	pohon dengan tinggi antara 10-20 meter, batangnya digunakan sebagai obat utk rematik dan sakit mata; [Bambusoideae]
8.	bangle	baŋlé	n	terna yang akarnya berwarna kuning, tinggi 1-1,5 meter, batangnya digunakan sebagai obat untuk radang; [Zingiber cassumunar]
9.	barora	barora	n	pohon, tinggi hingga 20 meter, ekstrak daunnya dapat digunakan sebagai obat utk sariawan; [Kleinhovia hospita]
10.	bawang merah	bawaŋ mérah	n	terna dengan tinggi 5 meter, biji dan daun dijadikan obat untuk eksem dan muntaber, [Allium cepa fa ascalonicum]
11.	bebele	bĕbèlè	n	tumbuhan berwujud tumbuhan menjalar, tinggi kurang dari 5 meter, daunnya untuk obat kencing nanah dan gondok; [Centella asiatica]
12.	belimbing wuluh	bĕlimbiŋ wuluh	n	pohon yang tingginya mencapai 15 meter, daunnya dapat digunakan sebagai obat untuk bisul dan rematik; [Averrhoa bilimbi]
13.	bengkel	bĕŋkĕl	n	pohon dengan tinggi mencapai 10 meter kulit batangnya digunakan untuk mengobati sakit ngilu dan layuh semu; [Nauclea Speciosa]
14.	beru	béru	n	waru; pohon kecil, tumbuh di sepanjang pantai, tetapi umum ditanam di pekarangan, bunganya berwarna kuning ketika pagi, sore hari berubah menjadi kemerah-merahan, kayunya banyak digunakan sebagai bahan bakar, daunnya dapat dijadikan obat bisul, [Hibiscus tiliaceus]
15.	bidara putih	bidara putih	n	pohon kecil tinggi bisa sampai 15 meter, terutama kulit akarnya, digunakan sebagai obat demam dan kayunya dapat digunakan untuk cuci darah [Strychnoss ligustrina]
16.	bila	bila	n	maja; pohon, tinggi mencapai 15 m, daunnya dijadikan obat eksem; [Aegle marmelos]
17.	birak	birak	n	eceng gondok; tumbuhan berupa tumbuhan air yang hidup terapung di permukaan air, tinggi kurang 5 meter, bunganya dapat digunakan sebagai obat biduran; [Eichhornia crassipes]
18.	bokah	bókah	n	labu air; tumbuhan berupa tanaman menjalar tinggi kurang 5 meter; daunnya dapat digunakan sebagai obat untuk bayi yang sakit panas; [Lagenaria lencantha]
19.	brotowali	brotowali	n	pohon dengan tinggi bisa mencapai 15 meter, rebusan batangnya dapat digunakan sebagai obat kolestrol dan gatal-gatal; [Tinospora Arispa]
20.	bunga sepatu	buŋa sĕpatu	n	perdu dengan tinggi 1-4 meter, bunga dapat dijadikan obat panas [Hibiscus rosasinensis]
21.	bunut	bunut	n	beringin; pohon besar yang tingginya mencapai 20–35 meter, daunnya dapat digunakan sebagai obat kejang karena panas dan disentri; [Ficus benjamina]
22.	cabai hutan	cabay hutan	n	tumbuhan berupa tanaman yang merambat, memanjat, membelit, dan melata seperti sirih, buahnya dapat menjadi obat masuk angin, beri-beri, dan pegal linu; [Piper retrofraktum]
23.	cemara	cĕmara	n	pohon yang berbatang tinggi lurus dengan tinggi 10—20 meter, daunnya dapat diekstrak untuk pengobatan penyakit kejang perut dan kulit bersisik; [Casuarina equisetifolia]
24.	dadap serep	dadap	n	pohon yg berukuran sedang, mencapai tinggi 15–20 meter, daunnya dapat digunakan untuk
		sèrèp		penyakit kencing batu; [Erythrina lithosperma]

25.	daun dewa	daun dèwa	n	terna dengan tinggi sekitar 30-40 cm, daunnya dapat digunakan untuk mengobati kolesterol; [Gynura divaricata]
26.	daun sendok	daun sèndok	n	terna yang tumbuh tegak, tinggi hingga 20 cm, akarnya dapat diolah untuk menjadi obat keputihan; [Plantago major]
27.	delima	dĕlima	n	perdu dengan tinggi 40—60 cm, buahnya dapat dijadikan obat sakit mata; [Punica granatum]
28.	entut-entut	ĕntut-ĕntut	n	tumbuhan berupa tanaman menjalar, daunnya digunakan untuk obat sakit perut dan sakit pinggang; [Paederia foetida]
29.	gandarusa	gandarusa	n	perdu yang tingginya 1–1,5 m, daunnya sebagai obat kencing batu; [Justicia gendarussa]
30.	gerepek	gĕrèpèk	n	pohon dengan tinggi 1 – 25 meter, daunnya dapat menjadi obat cacar; [Erythrina sp]
31.	inggu	iŋgu	n	terna tegak yang berdaun lebat, tinggi hingga 1,5 meter, daunnya untuk obat sakit kepala sebelah; [Ferula asa-foetida]
32.	jambu biji	jambu biji	n	perdu, dengan tinggi pohon dapat mencapai 9 meter; daunnya dijadikan obat sakit perut; [Psidium quajava]
33.	jarak pagar	jarak pagar	n	perdu besar dengan tinggi tanaman jarak pagar hanya sekitar 2 meter, daunnya dijadikan obat sakit perut; [Jatropha curcas]
34.	jeringo	jĕriŋó	n	jerangau;terna menahun dengan tinggi sekitar 75 cm, akarnya dapat digunakan sebagai bahan ramuan obat untuk penangkal makhluk halus; [Acorus calamus]
35.	jowet	jówét	n	jamblang; pohon, tingginya hingga 15 meter, buahnya dijadikan obat menceret; [Syzygium cumini]
36.	kacang hijau	kacaŋ hijau	n	perdu dengan tingginya 3 meter, batangnya bercabang tegak, bunganya dijadikan obat muntaber; [Vigna radiata]
37.	kanangas	kanaŋas	n	semak yang tingginya 2—4 meter, akarnya dijadikan obat eksem; [Ximenia sp]
38.	kangkung	kaŋkuŋ	n	tumbuhan berupa tanaman menjalar, daunnya dijadikan obat biduran; [Ipomoea reptans]
39.	kapas	kapas	n	semak dengan tinggi hingga 2 meter, buahnya mengandung serat berbulu putih yang digunakan untuk luka bakar; [Gossypium sp]
40.	kates	katès	n	pepaya; pohon setinggi 5–10 m , buahnya berdaging tebal dan manis untuk obat mencret; [Carica papaya]
41.	kecubung	kĕcubuŋ	n	perdu dengan ketinggian 3 meter, cabang-cabangnya berkayu, bunganya berbentuk corong dan berwarna putih atau ungu, bijinya digunakan obat panas dalam; [Datura fastuosa]
42.	kelapa	kĕlapa	n	pohon berbatang tinggi, buahnya dijadikan obat sariawan; [Cocos nucifera]
43.	kelor	kélór	n	pohon dengan tinggi hingga 8 meter, daun dijadikan menyehatkan badan; [Moringa oleifera]
44.	kemangi	kĕmaŋi	n	terna tinggi mencapai 150 cm, daunnya untuk obat sakit telinga; [Ocimum sanctum]
45.	kemuning	kĕmuniŋ	n	pohon renda tinggi mencapai 7 meter, daunnya dijadikan obat rematik;[Murraya paniculata]
46.	kenanga	kĕnaŋa	n	pohon, tinggi hingga 38 meter, bunganya dijadikan untuk obat panas; [Canangium odoratum]

47.	keroton	kerótón	n	perdu dengan tinggi 1-4 meter, bunganya dijadikan ramuan untuk membuat rambut tebal; [Hibiscus rosasinensis]
48.	kesambiq	kĕsambik	n	pohon, tinggi hingga 40 meter, daunnya dijadikan obat pusing tujuh hari; [Schleichera oleosa]
49.	kesembung	kĕsĕmbuŋ	n	sembung; perdu yang tumbuh tegak, tinggi 2–4 meter, daunnya dijadikan obat lemas; [Blumea balsamifera]
50.	ketimbus	kĕtimbus	n	ketimus; pohon tinggi yang batangnya untuk obat sakit koreng raja; [Protium javanicum burm]
51.	ketujur	kĕtujur	n	turi*;* pohon yang tingginya mencapai 12 meter, daunnya digunakan untuk melancarkan air susu ibu; [Sesbania grandiflora]
52.	ketumbar	kĕtumbar	n	perdu tinggi mencapai 1,3 meter, bijinya dijadikan obat panas; [Coriandrum sativum]
53.	klokos udang	klokos udaŋ	n	perdu tinggi yang buahnya dapat menjadi obat kulit bersisik; [Syzgium hemsiliana]
54.	kopi	kopi	n	perdu dengan tinggi 2-5 meter, bijinya dijadikan ramuan untuk menyegarkan kulit; [Coffea Arabica]
55.	kuluh	kuluh	n	keluih; pohon besar tinggi 10-25 meter, buahnya dijadikan ramuan untuk menyehatkan badan; [Artocarpus communis]
56.	kunyit	kuŋit	n	terna dengan tinggi dapat mencapai 100 cm, batangnya dapat dijadikan obat sakit perut; [Curcuma domestica]
57.	lada	lada	n	tumbuhan berupa tanaman memanjat dengan tinggi mencapai 15 meter, buahnya dapat dijadikan obat sakit perut; [Piper nigrum]
58.	laos	lawos	n	terna dengan tingginya 2 meter atau lebih, batangnya untuk obat bisul, panuan; [Alpinia galangal]
59.	lebui	lĕbuyi	n	gude; tanaman perdu dengan tinggi sekitar 0,5 – 4 meter, buahnya dijadikan ramuan untuk tambah darah; [Cajanus cajan]
60.	lekong	lĕkoŋ	n	pohon, tinggi hingga 39 meter, buahnya dijadikan ramuan untuk membuat rambut hitam; [Aleurites moluccana]
61.	lembain	lĕmbain	n	bayam; terna semusim daunnya dapat digunakan sebagai obat untuk penyakit biduran dan eksem, [Amaranthus sp]
62.	lita	lita	n	pulai; pohon pelindung , tingginya mencapai 10-50 meter, dapat dijadikan obat mag; [Alstonia scholaris]
63.	lobak	lobak	n	terna yang merupakan sayuran, buahnya dapat dijadikan ramuan untuk melancarkan persalinan; [Raphanus sativus]
64.	meniran	mĕniran	n	terna dengan tinggi hingga 1 meter, batangnya untuk obat ginjal; [Phyllanthus urinaria]
65.	merang	mĕraŋ	n	batang padi; terna semusim, batangnya untuk obat sakit perut; [Oriza sp.]
66.	nenas	nĕnas	n	terna dengan tinggi sekitar 20 sampai 30 cm, buahnya untuk ramuan melancarkan haid; [Ananas comosus]
67.	pakis	pakis	n	paku; perdu dengan tinggi kurang 5 meter, daunnya digunakan untuk obat biduran; [Cyathea/cycas sp.]
68.	pandan	pandan	n	pohon kecil yang tumbuh tegak hingga mencapai ketinggian 4–14 meter, daunnya dijadikan obat penyakit kulit; [Pandanu tectorius]
69.	paoq	paok	n	pohon yang berbatang tegak, daunnya dapat dijadikan obat silu; [Mangifera indica]

70.	pare	parè	n	paria; tumbuhan berupa tanaman menjalar, buahnya untuk obat sakit mata; [Momordica charantia]
71.	pecut kuda	pěcut kuda	n	terna yang tingginya mencapai 50 cm, daunnya dijadikan obat amandel; [Stachytarpheta jamaicensis]
72.	peko	péko	n	mengkudu; pohon mencapai 3-8 meter, daunnya digunakan sebagai obat sakit perut
73.	petai	pětay	n	pohon dapat mencapai 20 meter, buahnya dapat dijadikan obat borok; [Parkia speciosa]
74.	petikan kebo	pětikan kěbo	n	terna tegak dapat tumbuh hingga 60 cm, daunnya dapat menjadi obat untuk penyakit ginjal dan sakit mata; [Euphorbia hirta]
75.	pinang	pinaŋ	n	pohon dengan tinggi 25 meter, bijinya digunakan untuk penghilang bau mulut; [Area cathecu]
76.	raju mas	raju mas	n	pohon yang tinggi yang umumnya mencapai 45 meter, buahnya dijadikan obat diare; [Duabanga molucana]
77.	randu	randu	n	pohon tingginya tak lebih dari 2, daunnya dapat digunakan utk obat diare
78.	re	ré	n	alang-alang; semak yang tingginya kurang dari 5 meter, akarnya dijadikan obat kencing batu
79.	rembiga	rěmbiga	n	widuri; pohon kecil yang bergetah, tinggi hingga 3 meter, akarnya dapat dijadikan obat cacingan; [Calotropis gigantean]
80.	sager	sagěr	n	perdu, tinggi mencapai 3 meter, daunnya dapat dijadikan obat sariawan; [Sauropus androgynous]
81.	saladri	saladri	n	terna yang tingginya kurang dari 5 meter, daunnya dijadikan obat darah tinggi; [A. graveolens. L]
82.	salam	salam	n	pohonnya bertajuk lebat, tingginya mencapai 25 m, daunnya dijadikan obat sariawan; [Eugenia polyantha]
83.	sambiloto	sambiloto	n	terna dengan tinggi kurang 5 meter, daunnya dijadikan obat kencing manis; [Andrographis paniculata]
84.	sambung nyawa	sambuŋ ɲawa	n	terna dengan tinggi antara 30-45 cm, daunnya dijadikan obat pusing; [Gynura procumbens]
85.	sekuh	sěkuh	n	terna dengan tinggi 8 hingga 10 cm, batangnya dijadikan obat pegal linu; [Kaempferia galangal]
86.	selasih hitam	sělasih hitam	n	terna dengan tinggi 0,6—1,6 meter, bijinya dapat menjadi obat panas; [Ocimum basilicum L]
87.	semanggi	sěmaŋgi	n	tumbuhan berupa tanaman menjalar, daunnya dijadikan obat darah tinggi dan ginjal; [Hydrocotyle sibthorpioides]
88.	semet meong	sèmèt mèo ŋ	n	perdu dengan tinggi kurang 5 meter, daunnya direbus untuk obat penyakit ginjal; [Orthosiphon stamineus]
89.	sereto	sěrèto	n	perdu dengan tinggi 4-15 meter, batangnya dapat dijadikan obat malaria; [Ehretia microphyla]
90.	serikaya	sěrikaya	n	tanaman perdu yang tingginya mencapai 2–7 meter, buahnya dapat dijadikan obat malaria; [Anona squamosal]
91.	sirih	sirih	n	tumbuhan berupa tanaman merambat di pohon lain, daunnya digunakan untuk ramuan penghilang bau mulut; [Piper betle]
92.	soka	soka	n	pohon yang tingginya bisa mencapai lebih dari 4 meter, batang dan akarnya bisa direbus untuk obat panas dan luka; [Ixora paludosa]

93.	songgo langit	soŋgo laŋit	n	semak dengan tinggi di bawah 5 meter, daunnya dijadikan obat ginjal; [Tridax procumbens]
94.	suren	surèn	n	surian; pohon dengan tinggi mencapai 35 meter, batangnya dijadikan obat bisul; [Toona sureni]
95.	tapak dara	tapak dara	n	perdu dengan tinggi kurang dari 5 meter, daunnya dijadikan untuk obat kencing manis; [Catharanthus receous]
96.	tebu besi	tĕbu bĕsi	n	semak dengan tinggi mulai dari 2,5 meter hingga 4 meter, batangnya dapat menjadi ramuan antiuban; [Saccharum officinarum]
97.	tembakau	tĕmbakaw	n	pohon yang tumbuh hingga ketinggian antara 1 sampai 2 meter, daunnya digunakan untuk obat borok; [Nicotiana tabacum]
98.	temulawak	tĕmulawak	n	terna, tinggi hingga 2,5 meter, irisan rimpang yang dikeringkan dapat menjadi obat flu dan demam; [Curcuma xanthorrhiza]
99.	terung	tĕruŋ	n	perdu dengan tinggi tanaman sekitar 3 meter, buahnya dijadikan obat rematik; [Solanum melongena]
100.	wareng	warèŋ	n	pohon dengan tinggi rata-rata 21,67 meter, buahnya dijadikan obat gusi berdarah; [Gmelina elliptica]
101.	waru laut	waru laut	n	pohon dengan tinggi 2–10 meter, daunnya dijadikan obat silu; [Thespesia populnea]
102.	wortel	wortĕl	n	terna tumbuh antara 30 dan 60 cm, buahnya dijadikan obat desentri; [Daucus carota]

LANGUAGE DOCUMENTATION ON MBOJO'S TRADITIONAL AGRICULTURAL TOOLS

Nuryati
Regional Office for Language in West Nusa Tenggara Barat Province, Indonesia
nuryati@kemdikbud.go.id

Abstract

Every ethnicity in Indonesia has its own characteristics and uniqueness that becomes their identity and has the values of local wisdom. These characteristics and uniqueness can be in the form of language, customs, clothes, traditional food, and so on. West Nusa Tenggara is one of the provinces that is unique in the existence of three indigenous ethnicities, namely Sasak, Samawa, and Mbojo. Most of West Nusa Tenggara is a fertile area so that many people from ancient times lived from agriculture. Along with the times and technology advances as it is today, agriculture with traditional models has been largely abandoned and replaced with modern equipment, for example the existence of plows has been replaced by tractor engines. The replacement of this plow certainly means that community and the younger generation will no longer know what a plow is. The plow as one of the agricultural tools driven by cows or buffaloes has many parts with their respective terms.This shows that the existence of this tractor engine is able to drown the a lot of vocabulary or terms in a plow, such as the term part of the cow's place, parts of plow wood, and so on. The language documentation related to plows as a traditional agricultural tool is an important thing to do so that people, especially the younger generation, can still recognize it because actually the vocabulary in a language is a historical track record of the life of an ethnic community.

Key words: Mbojo ethnicity, agriculture, plow.

Pendahuluan

Bangsa Indonesia terdiri dari berbagai suku bangsa atau etnis. Dari sekian etnis atau suku bangsa tersebut, terdapat tiga etnis besar di Nusa Tenggara Barat yang sampai hari ini masih hidup dengan bahasa dan budaya mereka. Ketiga etnis tersebut adalah etnis Sasak di Pulau Lombok, etnis Samawa di Pulau Sumbawa sebelah Barat, dan etnis Mbojo di Pulau Sumbawa sebelah Timur.

Sebagaimana etnis-etnis atau suku-suku bangsa yang lain, ketiga etnis di Nusa Tenggara Barat ini juga memiliki keunikan budaya dan tradisi, seperti budaya atau tradisi-tradisi agraris, maritim, dan siklus kehidupan, yang semua itu telah memperkaya khazanah budaya Nusa Tenggara Barat (Taufan, 2012).

Bercocok tanam atau bertani adalah mata pencaharian utama bagi sekitar delapan puluh masyarakat Nusa Tenggara Barat. Perubahan tata cara kehidupan dari masyarakat berburu menjadi masyarakat yang bercocok tanam merupakan pertanda awal dari cara berpikir dan sikap masyarakat dari yang bersahaja menuju cara berpikir yang lebih maju dan memerlukan keterampilan. Menurut catatan sejarah, pulau Sumbawa terutama daerah Bima dan Dompu mengenal pertanian dengan sistem persawahan itu sekitar abad XIV. Seorang Raja Bima yang bernama Ruma Nawaa Paju Langge mengirim saudaranya ke Goa di kerajaan Manurung untuk belajar berbagai ilmu dan pengetahuan diantaranya tentang pertanian persawahan dan pengairan. Sejak itu masyarakat Bima dan Dompu (Mbojo) mengubah tata cara bercocok tanam dari sistem berhuma ke sistem persawahan dan pengairan. Alat-alat pertanian seperti luku atau bajak dan peranan kerbau sebagai penarik bajak mulai dikenal dan dipergunakan (Proyek Pengembangan Permuseuman Propinsi NTB, 1982)

Pertanian yang dilakukan secara tradisional oleh masyarakat telah mengalami perkembangan dengan cara bertani yang modern mengikuti perkembangan teknologi. Menurut Kamus Besar Bahasa Indonesia, pertanian berasal dari kata *tani* yang artinya 'mata pencaharian dalam bentuk bercocok tanam; mata pencaharian dalam bentuk mengusahakan tanah dengan tanam-menanam. Sementara, pertanian itu sendiri adalah perihal bertani (mengusahakan tanah dengan tanam-menanam) (KBBI, 2015:1400).

Secara umum, pengertian dari pertanian adalah suatu kegiatan manusia yang termasuk di dalamnya yaitu bercocok tanam, peternakan, perikanan, dan juga kehutanan. Sebagian besar mata pencaharian masyarakat di Indonesia adalah sebagai petani, sehingga sektor pertanian sangat penting untuk dikembangkan di negara kita.

Ada beberapa bentuk-bentuk pertanian di Indonesia: (1) Sawah; sawah adalah suatu bentuk pertanian yang dilakukan di lahan basah dan memerlukan banyak air baik sawah irigasi, sawah lebak, sawah tadah hujan, maupun sawah pasang surut. (2) Tegalan; tegalan adalah suatu daerah dengan lahan kering yang bergantung pada pengairan air hujan, ditanami tanaman musiman atau tahunan dan terpisah dari lingkungan dalam sekitar rumah. Lahan tegalan tanahnya sulit untuk dibuat pengairan irigasi karena permukaan yang tidak rata. Pada saat musim kemarau lahan tegalan akan kering dan sulit untuk dibubuhi tanaman pertanian. (3) Pekarangan; pekarangan adalah suatu lahan yang berada di lingkungan dalam rumah (biasanya dipagari dan masuk ke wilayah rumah) yang dimanfaatkan/digunakan untuk ditanami tanaman pertanian. (4) Ladang Berpindah; ladang berpindah adalah suatu kegiatan pertanian yang dilakukan di banyak lahan hasil pembukaan hutan atau semak yang setelah beberapa kali panen/ditanami, maka tanah sudah tidak subur sehingga perlu pindah ke lahan lain yang subur atau lahan yang sudah lama tidak digarap.

Seiring perkembangan zaman dan kemajuan teknologi sistem pertanian tradisional telah mulai bergeser dengan diciptakannya alat-alat pertanian yang lebih mengedepankan teknologi modern, seperti munculnya traktor sebagai pengganti alat bajak tradisional. Kemunculan dan perkembangan penggunaan alat tersebut tentu sangat menguntungkan para petani karena dengan menggunakan alat tersebut pekerjaan dapat diselesaikan dengan lebih cepat dan lebih praktis. Perkembangan zaman dan kemajuan teknologi ini dapat dipandang bagai dua sisi mata uang dari sisi kebahasaan. Kemunculan sebuah teknologi baru mampu memberikan tambahan kosakata baru dan juga sebaliknya, mampu menenggelamkan kosakata yang ada. Hal ini dapat kita ambil contoh dengan munculnya teknologi komputer pada waktu itu tentu akan mengiringi kemunculan kosakata-kosakata baru terkait perangkat dan sistem dalam komputer tersebut yang manjadi penambah kekayaan kosakata. Namun, hal sebaliknya juga kita bisa perhatikan bahwa kemunculan alat seperti traktor dalam pertanian ini tentu akan menenggelamkan kosakata-kosakata pada alat pertanian yang digunakan sebelumnya yaitu bajak. Keberadaan sebuah kosakata dalam suatu bahasa merupakan rekam jejak keberlangsungan kehidupan suatu etnis tertentu yang syarat dengan nilai-nilai kearifan lokal. Seiring dengan kemunculan teknologi baru yang menenggelamkan keberadaan kosakata-kosakata pada teknologi yang ada sebelumnya ini haruslah dibarengi pula dengan usaha nyata sebagai bentuk pendokumentasian kosakata dalam suatu bahasa yang terancam tenggelam tersebut. Pendokumentasian bahasa melalui bidang leksikografi dapat diawali dengan inventarisasi kosakata yang nantinya dapat disajikan dalam bentuk kamus, tesaurus, glosarium atau ensiklopedia.

Metode

Kegiatan pendokumentasian bahasa ini adalah suatu kegiatan penelitian yang dibagi dalam tiga tahapan strategis, yaitu tahap prapenelitian, tahap penelitian, dan tahap pascapenelitian. Tahap prapenelitian mencakup kegiatan penentuan tim pelaksana, pengumpulan informasi, penyusunan proposal, dan penyusunan instrumen penelitian. Tahap penelitian mencakup kegiatan pengumpulan data, analisis data, dan penyusunan hasil analisis data. Adapun tahap pascapenelitian mencakup kegiatan penyusunan laporan penelitian.

Dalam kegiatan metode yang digunakan adalah metode lapangan dengan teknik wawancara. Sementara itu, populasi penelitian ini adalah seluruh kosakata bahasa Mbojo yang terkait dengan alat pertanian khususnya bajak yang dipakai oleh penutur asli bahasa Mbojo baik berdasarkan geografi dialek maupun geografi sosial. Jumlah informan secara keseluruhan ialah sebanyak 30 orang dengan kriteria sebagai berikut:

1. Informan merupakan penutur asli bahasa yang diteliti.
2. Informan berusia 35 tahun ke atas.
3. Informan mempunyai intelegensi yang cukup tinggi dan setidak-tidaknya berpendidikan SLTP.
4. Informan tidak terlalu lama meninggalkan tempat asal.
5. Informan dapat berbahasa Indonesia.
6. Informan tidak cacat wicara.
7. Informan tidak terlalu lama menggunakan bahasa lain secara terus-menerus.
8. Informan bersedia menjadi informan.
9. Informan bersikap terbuka, ramah, jujur, dan tidak terlalu emosional dan mudah tersinggung.
10. Informan memiliki daya ingatan yang baik, tidak pemalu dan suka berbicara. (Taryono dalam Susilo 1998: 6)

Instrumen

Instrumen yang digunakan dalam penelitian pendokumentasian kosakata alat pertanian Mbojo ini adalah bagian-bagian yang berkaitan dengat alat bajak pada pertanian Mbojo.

Alat Pertanian Bajak		
No.	**Kosakata dalam bahasa Mbojo**	**Definisi/Keterangan**
1.		
2.		
3.		
4.		
dst.		

Data dan Teknik Pengumpulan Data

Data dalam kegiatan pendokumentasian ini adalah kosakata yang berkaitan dengan alat pertanian Mbojo khususnya alat bajaknya.

Pengumpulan data dilakukan melalui studi data di lapangan. Studi lapangan tersebut dengan menggunakan metode cakap (Mahsun, 2005). Adapun metode cakap dengan teknik catat, peneliti dapat langsung mencatat hal-hal yang berhubungan dengan kosakata bahasa tersebut melalui wawancara. Di samping itu peneliti juga melibatkan diri sebagai informan/penyedia data (lihat Mahsun 2003: 85). Pengumpulan data juga dilakukan dengan studi pustaka terkait alat pertanian Mbojo.

Teknik Analisis Data

Data yang sudah terkumpul akan dianalisis dengan menggunakan metode padan ekstralingual (Mahsun 2003:114) kemudian dilanjutkan dengan metode deskriptif kualitatif teknik *content analysis,* yaitu suatu teknik analisis yang digunakan untuk menerjemahkan secara sistematis dan objektif berbagai pesan dan pernyataan yang diperoleh dari wawancara mendalam dengan informan (Berg, 1989).

Teknik Penyajian Hasil Analisis Data

Hasil analisis data akan disajikan dengan teknik formal dan informal seperti yang disarankan (Sudaryanto dalam Mahsun, 2005). Yang dimaksud dengan teknik informal adalah perumusan dengan kata-kata biasa walaupun dengan terminologi yang bersifat teknis. Karena penyusunan glosarium ini mendeskripsikan bentuk-bentuk leksikal, maka lema yang menjadi padanan lema dan sublema pada lema umum (bahasa standar atau bahasa jamak). Hal ini dilakukan untuk memudahkan pembaca untuk dapat langsung membedakan mana bentuk-bentuk yang menjadi padanan kata-kata tersebut.

Data yang sudah terseleksi pada langkah analisis data kemudian akan dilanjutkan dengan langkah berikutnya yaitu teknik penyajian data. Teknik penyajian data ini mencakup dua bidang hal yaitu pengabjadan dan pemberian definisi. Pengabjadan merupakan pekerjaan yang memerlukan ketekunan, ketelitian, kesabaran, dan kepatuhan terhadap kesepakatan bersama yang berkaitan dengan teknik leksikografis serta patokan-patokan khusus yang disepakati bersama untuk dilaksanakan. Kesepakatan ini harus dipatuhi agar hasil akhir penyusunan akan menampakkan keteraturan. Pengabjadan tersebut dilakukan secara horisontal dan vertikal. Yang dimaksud dengan pendefinisian adalah salah satu kegiatan dalam penyusunan kamus yang memerlukan ketenangan, ketekunan, ketajaman analisis, ketelitian, kecermatan, kesabaran, dan wawasan yang luas. Kesalahan dalam memberikan batasan makna kata berarti menjerumuskan pemakai kamus (Sunaryo, 1984)

Dalam kegiatan leksikografi secara umum mempunyai langkah-langkah yang hampir sama yaitu (a) Penentuan sumber data, (b) teknik pengaturan data, (c) teknik penyeleksian data, (d) teknik penyajian data, (e) teknik penyusunan entri, (f) teknik pengetikan naskah, dan (g) lambang ortografi.

a. Penentuan sumber data
   Sumber data untuk kegiatan leksikografi adalah bahasa. Penentuan sumber data ini menjadi langkah awal dalam rangkaian kegiatan dalam leksikografi. Sumber data ini bisa diperoleh dari kajian pustaka dan pengambilan data di lapangan.

b. Pengaturan data
   Data masukan yang baik dan berguna apabila ditangani secara baik dan bersistem sesuai dengan tujuan penyusunan yang akan dilaksanakan.

c. Teknik penyeleksian data
   Data yang telah diperoleh dari lapangan kemudian akan diolah dengan langkah-langkah yaitu (1) data dilihat secara keseluruhan dan memisahkan data yang diperlukan dari data yang harus disisihkan, (2) data dikelompokkan berdasarkan bentuk kata atau entri masukan), (3) data dipilah berdasarkan medan makna, dan (4) data disusun menurut abjad perkelompok bentuk entri masukan.

d. Teknik penyajian data
   Data yang sudah terseleksi pada langkah analisis data kemudian akan dilanjutkan dengan langkah berikutnya yaitu teknik penyajian data. Teknik penyajian data ini mencakup dua bidang hal yaitu pengabjadan dan pemberian definisi. Pengabjadan merupakan pekerjaan yang memerlukan ketekunan, ketelitian, kesabaran, dan kepatuhan terhadap kesepakatan bersama yang berkaitan dengan teknik leksikografis serta patokan-patokan khusus yang disepakati bersama untuk dilaksanakan. Kesepakatan ini harus dipatuhi agar hasil akhir penyusunan akan menampakkan keteraturan. Pengabjadan tersebut dilakukan secara horisontal dan vertikal. Yang dimaksud dengan pendefinisian adalah penjelasan makna kata yang harus memperhatikan kesejajaran antara lema dengan makna yang diberikan.

   Definisi terdiri atas beberapa jenis yaitu:

   - Definisi leksikografis yaitu mendekripsikan secara berurutan ciri-ciri semantik terpenting, umumnya berupa penjelasan singkat dan sederhana, contoh:
     manusia *n* mahkluk berakal dan berbudi (dibedakan dari binatang)

- Definisi sinonimis yaitu berupa padanan kata yang sama atau mirip (dalam kamus ekabahasa definisi dapat digunakan melengkapi definisi leksikografis), contoh:
  manusia *n* insan;orang
- Definisi logis yaitu secara tegas mengidentifikasikan obyek yang dideskripsikan sehingga membedakannya dari obyek lain dan menggolongkan secara tegas sebagai anggota golongan yang terekat (lebih bersifat ilmiah); definisi ini biasanya digunakan dalam kamus bidang ilmu, contoh:
  manusia *n* mahkluk yang berakal dan berbudaya, daif, dan fana (dibedakan dari binatang dan malaikat)
- Definisi ensiklopedis yaitu memberikan gambaran secara lengkap dan cermat segala sesuatu yang berhubungan dengan entri yang diberi definisi, contoh:
  air *n* persenyawaan hidrogen dan oksigen, terdapat di mana-mana dan dapat berwujud gas (uap air), cairan, dan zat padat (es atau salju); air adalah zat pelarut yg baik sekali terdapat di alam di keadaan tidak murni; air murni berupa cairan yg tidak berbau.......dst

e. Teknik penyusunan entri
Dalam penyusunan entri diperlukan kekonsistenan dan ketaatasasan agar hasil akhir yang diperoleh dapat tersusun sistematis. Penulisan entri utama dan entri-entri tambahan yang mengikutinya harus didasarkan pada kesepakatan bersama. Pengabjadan lema dilakukan secara vertikal dan horizontal.

f. Teknik pengetikan naskah
Langkah yang dilakukan keseluruhan pengolahan data yaitu pengetikan hasil akhir yang sudah dilengkapi dengan pemberian definisi.

g. Lambang ortografi
Suatu bahasa mempunyai sistem tulisan standar dan sistematis. Penggunaan lambang ortografis berguna untuk memberikan informasi singkat yang mengenai sasaran tentang sebuah lema.

Hasil Data dan Pembahasan

Usaha pendokumentasian alat pertanian tradisional Mbojo ini dikhususkan pada alat bajak yang keberadaanya sekarang telah mulai tergeserkan dengan keberadaan sebuah alat menggunakan teknologi yaitu traktor. Alat bajak tradisional dan alat traktor ini mempunyai kesamaan fungsi untuk mengolah tanah agar siap untuk ditanami. Kesamaan bagian yang terdapat pada dalam kedua alat itu adalah pada mata bajak. Namun demikian, tahapan penggunaannya juga berbeda antara kedua alat tersebut. Penggunaan alat bajak modern ini jauh lebih menghemat waktu karena kerjanya lebih cepat dibandingkan dengan alat bajak tradisional. Alat bajak tradisional yang masih menggunakan tenaga kerbau untuk masyarakat Mbojo ini tentu pengerjaannya menghabiskan waktu yang lebih lama untuk proses membajak dari tahap awal sampai terakhir hingga tanah siap untuk ditanami. Dengan tergesernya alat bajak tradisional ini tentu juga akan menenggelamkan kosakata-kosakata yang menunjukkan nama bagian-bagian dari alat tersebut. Semakin tergesernya keberadaan kosakata-kosakata tersebut tentunya akan semakin hilang penggunaannya oleh masyarakat terutama generasi muda yang lebih mengetahui keberadaan sebuah traktor daripada alat bajak tradisional tersebut.

Hasil data alat pertanian tradisional Mbojo khususnya alat bajak ini yang dapat dikumpulkan sebagai bahan dokumentasi bahasa adalah sebagai berikut.

Alat Pertanian Bajak		
No.	Kosakata dalam bahasa Mbojo	Definisi/Keterangan
1.	*ai pehi*	tali kendali yang terbuat dari rotan atau kulit kerbau berfungsi untuk menghubungkan atau menyatukan antara *nggala* dengan *oka*.
2.	*cambo*	tali pecut
3.	*cau*	alat menyerupai sisir yang terbuat dari batang kayu panjangnya sekitar 1 m tebal 15 cm yang digunakan untuk menyisir tanah
4.	*cau hade*	proses keempat menyisir atau meratakan tahap akhir
5.	*cau rea*	proses ketiga menyisir atau meratakan tanah tahap awal
6.	*dou marawi*	orang yang mengendalikan bajak
7.	*garanci asa sahe*	penutup mulut kerbau
8.	*geligi cau*	
9.	*jee jee*	nyanyian yang dilakukan oleh pengendali bajak untuk membantu mengarahkan kerbau
10.	*karaci*	tali pecut
11.	*karepa sahe*	kalung kerbau
12.	*leto nggala*	berfungsi sebagai tempat pegangan pengendali *nggala*/bajak
13.	*nao*	bagian besi gigi yang terletak pada bagian bawah yang berdekatan dengan tanah pada lekukan kayu *nggala*
14.	*nggaka*	proses pertama mambajk sawah yang masih utuh
15.	*nggala*	komponan bagian belakang bajak yang terbuat dari sepotong kayu yang agak bulat panjangnya antara 75—100 cm berbentuk seperti huruf L.
16.	*oka*	bagian alat yang diletakkan pada leher hewan penarik bajak (kerbau). Bagian ini berfungsi untuk menyatukan hewan penarik dengan bajak tersebut dengan cara mengapit leher kerbau yang dengan kayu yang terpasang pada kedua sisi *oka*. Bagian ini dibuat dari kayu dengan panjang sekitar 100 cm tebal 15 cm
17.	*rawi*	proses membajak
18.	*sahe*	kerbau
19.	*santira cau*	alat yang terbuat dari kayu bercabang pada salah satu ujungnya dan dibuatkan alat untuk tempat penghubung dengat alat *cau*. Pada sambungan itu dapat digerakkan sesuai yang diinginkan oleh pengendali *cau* baik mengubah arah atau melepas kotoran yang tersangkut pada *geligi cau*.

20.	*santira nggala*	bagian bajak yang dibuat dari sebatang kayu panjangnya 250 cm, tebal sekitar 14 cm dipasang untuk menyambung pada *nggala*. Ini merupakan poros penghubung antara bagian bajak depan dan belakang. Ujung depan *santira nggala* ini merupakan titik tengah *oka* yang merupakan titik tumpu yang dapat mengatur jalannya kerbau sesuai dengan arah yang ditentukan oleh pengendali
21.		proses membajak kedua dengan memotong arah dari bekas hasil bajakan pertama (*nggaka*)

Dari satu alat pertanian tradisional bajak pada etnis Mbojo ini telah menunjukkan kosakata-kosakata yang menjadi bagian-bagian dari sebuah bajak. Usaha pendokumentasian ini diharapkan bisa bermanfaat untuk memberikan gambaran tentang alat pertanian tradisional Mbojo khususnya alat bajak yang dulunya dipakai oleh masyarakat Mbojo.

Daftar Pustaka

Nuryati., Hartini., Hariro, Z., Cahyasabudhi, I. N. (2018). *Laporan Kegiatan Penyusunan Glosarium Bahasa Samawa Bidang Pertanian.* Mataram: Kantor Bahasa Provinsi Nusa Tenggara Barat.

Mahsun. (2005). *Metode Penelitian Bahasa: Tahapan Strategi, Metode, dan Tekniknya*. Jakarta: PT Raja Grafindo Persada.

Pengembangan Permuseuman NTB.(1982). *Alat Pertanian dan Fungsinya di Propinsi Nusa Tenggara Barat*. Mataram. Museum Propinsi Nusa Tenggara Barat

Sunaryo, A., dkk. (1990). *Pedoman Penyusunan Kamus Dwibahasa*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Taufan, N.I. (2012). *Warna-warni Tradisi Sasak Samawa Mbojo*. Bima: Museum Kebudayaan Samparaja

ASIAN ENGLISHES AND LEXICOGRAPHY

Danica Salazar
Oxford English Dictionary
danica.salazar@oup.com

Abstract

English has been present in Asia for hundreds of years, and throughout this time, the English lexicon has been profoundly influenced by other Asian languages, and significantly shaped by innovations made by its millions of Asian speakers. This talk will give a chronological account of various efforts to document the distinctive word store of Asian Englishes—from the earliest wordlists, glossaries, and dictionaries compiled by Western explorers, settlers, and colonizers in Asia, to the coverage of Asian words in large general dictionaries, particularly the historical Oxford English Dictionary, to smaller localized dictionaries written by lexicographers from or based in Asia. As will be shown in the talk, these lexicographical projects reflect changing user attitudes towards nativized vocabulary, the evolving theoretical perspectives through which these words have been viewed by linguists, philologists, and lexicographers, and the various roles that English has played in Asia over the centuries. They also demonstrate how Asian contributions to the English lexicon have developed from being mostly exotic loanwords borrowed from local vernaculars to more varied and creative constructions that illustrate the hybrid and translingual nature of the everyday lexis of contemporary Asians.

Keywords: Asian Englishes, history of lexicography, Oxford English Dictionary

CLIMATE CHANGE VS GLOBAL WARMING: A CORPUS-DRIVEN APPROACH TO CLIMATE TALK OF THE DECADE 2010-2019

Locky Law
Centre for Applied English Studies, the University of Hong Kong
lockylaw@hku.hk

Abstract

Ecolinguistics, a strand of research pioneered by Halliday (1990) focusing on the impact of language on the environment, has traditionally taken an approach similar to critical discourse analysis (CDA) (Baker & Ellece, 2011; Fill & Muhlhausler, 2001). In the last decade, the application of corpus-assisted methods in ecolinguistics studies began to draw attention. This presentation extends the talk by Law and Matthiessen (2019) with an aim to look deeper into the lexicographical patterns in climate talk between 2010 and 2019. A 227,499-word eco-corpus was created from a collection of 275 randomly selected texts on the topic of environment published on the internet within this period. These texts consist of media reports, online magazines, transcripts from TV talk shows, and public speeches. A corpus linguistics analysis emphasizing on 1-gram, 2-word concgrams, and 3-word concgrams and their respective concordances was performed using ConcGram 1.0 (Greaves, 2009). Findings revealed the top 40 unique words, unique nouns, meaningful 2-word and 3-word concgrams, as well as distinctive trends in word choices. This includes the preference for *climate change over global warming,* the low frequencies of occurrence for words related to wildlife (e.g., whale, bee, frog, fish, bird, reef), and the more frequent use of neutral and negative words (e.g. risk, cause, issues) than positive ones (e.g. reafforestation, healthy, promote, ratify). The concgrams were also grouped by cause, effect, time and places/people. These results provide a snapshot of the language use by the media outlets in construing climate talk in the 2010s. Alternative words to approach climate talk in the 2020s will be discussed.

Keywords: ecolinguistics, climate talk, climate change, global warming, concgram

References

Baker, P. & Ellece, S. (2011). *Key Terms in Discourse Analysis*. Bloomsbury UK.

Fill, A. & Muhlhausler, P. (2001). *The ecolinguistics reader: Language, ecology, and environment*. Continuum.

Greaves, C. (2009, February 5). *ConcGram 1.0: A phraseological search engine - user manual*. John Benjamins Publishing. http://www.benjamins.eom/jbp/series/CLS/1/manual.pdf

Halliday, M. AK. (1990). New ways of analysing meaning: A challenge to applied linguistics. *Journal of Applied Linguistics*, *6*, 7-36.

Law, L. & Matthiessen, C. MIM. (2019, September 5-7). *Revisiting Halliday's (1990) "New ways of meaning: The challenge to applied linguistics": What has changed and what still needs to be done?* [Paper presentation]. The Conference on Language and Ecology: Towards a Shared Narrative in Interdisciplinary Research 2019, Hong Kong. https://www.ecolinguistics2019.com/

DEVELOPMENT OF EFL DICTIONARIES WITH SPECIAL REFERENCE TO THEIR INNOVATIVE FEATURES

Shigeru Yamada
Waseda University, Japan
shyamada@waseda.jp

Abstract

EFL dictionaries have been developing, bringing in special, innovative features which set the dictionaries apart from other genres. The defining vocabulary and meticulous verb patterns were introduced by West's New Method English Dictionary (1935) and Palmer's A Grammar of English Words (1938), respectively. The first full-fledged EFL dictionary, Idiomatic Syntactic English Dictionary (1942), concentrated on contemporary lexical information and provided it in simple language, using the International Phonetic Alphabet, clearly indicating the countability and uncountablity of nouns, and providing ample collocations and examples. The first corpus-based dictionary, Collins COBUILD English Language Dictionary (1987), has revolutionized dictionary compilation. It depended on frequency information for choice of headword items, identification and arrangement of senses, and selection and ordering of collocations and examples. The dictionary also used full-sentence definitions universally. Its 2nd edition (1995) indicated the relative importance of headwords by means of the five-level Frequency Bands. The 3rd edition of Longman Dictionary of Contemporary English (1995) incorporated menus and signposts to make it easier for the user to navigate long, polysemous entries. This paper looks at the development of EFL dictionaries, paying special attention to their main innovations: how they were inspired, what significance and impact they have had, and how they were adopted and adapted by rival and other dictionaries.

Keywords: corpus, defining vocabulary, EFL dictionaries, full-sentence definition, grading of headwords, indication of verb patterns, menus, signposts

THE LEXICAL IMPACT OF COVID-19 AND ITS LEXICOGRAPHICAL DESCRIPTION

Yongwei Gao
Fudan University, China
ywgao@fudan.edu.cn

Abstract

COVID-19 has wrought great havoc worldwide since the beginning of 2020. For most people, life has been brought to a sudden standstill or disrupted to a great extent as the plague impacts every aspect of human endeavor. As the world fights against the novel coronavirus, many words and expressions associated with the pandemic have been created. These terms range from technical terms to everyday words related to daily life. Some of them concern themselves with the symptoms or medical conditions of the disease while others have something to do with the policies adopted by national governments; some are used specifically by the medical profession and researchers while others may be uttered by the man on the streets. As the number of such terms increases, dictionary-makers worldwide have gone out of their way to record them in a timely fashion. The Oxford English Dictionary, The Merriam-Webster Collegiate Dictionary and even Oxford Advanced Learner's Dictionary provided an update on COVID-related terms. The OED, for instance, recorded some twenty neologisms in its April update that include infodemic, social distancing, to flatten the curve, WFH, etc. In its July update, the OED added more than 40 COVID-related neologisms that range from corona to spike protein. Efforts have also been made to compile dictionaries that focus solely on words and expressions closely related to the pandemic. As is evidenced in the frequency in COVID-related terms in news reports or corpora such as NOW (News on the Web), it is no exaggeration to say that the lexical impact exerted by COVID-19 is greater than any pandemic in human history. This paper attempts to make an in-depth discussion of the lexical impact of the on-going pandemic and review the lexicographical effort in recording such impact.

Keywords: COVID-19, neologisms, terms, dictionaries

ON THE NEED OF BILINGUALISED CHINESE-ENGLISH DICTIONARIES WITH SPECIAL REFERENCE TO *ZHONG FC*

Yu-kit Cheung
Department of Translation, Lingnan University, Hong Kong
yukitcheung@LN.edu.hk

Abstract

Monodirectional Chinese-English dictionaries are often turned to in tasks of Chinese- English translation by native speakers of Chinese, for as Adamska-Salaciak (2016) remarks, 'the primary task of an L1-L2 dictionary is to serve as an aid in the user's own foreign language production' (p. 145). To the dictionary makers of such works as in other L1-L2 dictionaries in general, understanding of the L1 lemma in question is taken for granted (ibid, p. 144).

Nevertheless, this assumption seems to be put under challenge as far as Chinese being L1 is concerned. With reference to the English equivalents of '*zhong*&' in a wide array of Chinese-English dictionaries such as Du (2016), Lin (1971), Mathew (1943), to name but three, it will be argued that imprecision or errors in Chinese-English translation are on occasions attributed to misreading of the Chinese source text on the part of the user on the one hand, and omission of individual senses or misinterpretation of the headword on the part of the Chinese lexicographer on the other.

It is posited that the problem at issue arises from the concurrency of two sets of vocabulary in Standard Modern Chinese - lexical items from both Classical Chinese (*Wenyan* ^･) and Vernacular Chinese (*Baihua gg)* and from the assumption that Chinese-English dictionary users, who are by and large native speakers of Chinese, have the natural endowment to make an educated choice out of a collection of suggested equivalents. It is, therefore, proposed to explore the feasibility of partial bilingualised dictionaries, if not fully bilingualised ones, in contributing to enhancing the accuracy of English translation.

This paper is significant in challenging the established assumption of a bilingual dictionary user's firm grasp of the source language, throwing light on the need of bilingualised Chinese-English dictionaries in Chinese-English translation and pushing back the frontiers of bilingual lexicography.

Keywords: Chinese-English translation, Chinese English dictionaries, bilingualisation

References

Adamska-Salaciak, A. (2016). Explaining meaning in bilingual dictionaries. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (144-160). Oxford: Oxford University Press.

Du, R.Q. (2016). *New century Chinese-English dictionary*. Beijing: Foreign Language Teaching and Research press.

Lin, Y. (1972). *Lin Yutang's Chinese-English dictionary of modern usage*. Hong Kong: The Chinese University of Hong Kong.

Mathews, R. H. (1943). Mathews's Chinese-English dictionary. (Rev. Am. ed..). Harvard University Press.

National Agency for Language Development and Cultivation
Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia

Jalan Daksinapati Barat IV, Rawamangun
Jakarta Timur 13220
Telepon (021) 4706287, 4706288, 4896558, 4894564
http://badanbahasa.kemdikbud.go.id

9 786233 071345