



ASIALEX 2019 İSTANBUL

PROCEEDINGS of the 13th International Conference of the Asian Association for Lexicography

Editors

**Mehmet Gürlek, Ph.D.
Ahmet Naim Çiçekler, Ph.D.
Yasin Taşdemir**



asialex2019



asialex2019

<http://asialex2019.istanbul.edu.tr>
asialex2019@gmail.com

Istanbul University Department of Linguistics, Istanbul-Turkey

The Asian Association for Lexicography (ASIALEX) Proceedings Book

ISBN: 978-605-7736-04-8

Editors

Mehmet Gürlek

Ahmet Naim Çiçekler

Yasin Taşdemir

Cover Design

Bülent Polat

Date of Publishing

09.09.2019

Asos Publisher

1. Print

Address: Çaydaçıra Neighborhood Hacı Ömer Bilginoğlu Street No: 67 / 2-4 / MERKEZ / ELAZIĞ

Telephone: 0532 643 75 23

Mail: asos@asosyayinlari.com

Web: www.asosyayinlari.com

Instagram: https://www.instagram.com/asosyayinevi/

Facebook: https://www.facebook.com/asosyayinevi/

Twitter: https://twitter.com/Asosyayinevi

BOARD

Rachel Edita O. ROXAS

President

National University

(The Philippines)

Vincent B Y OOI

Vice-President

National University of Singapore

(Singapore)

Shirley DITA

Secretary

De La Salle University

(The Philippines)

Deny KWARY

Treasurer

Universitas Airlangga

(Indonesia)

Yongwei GAO

Board member

Fudan University

(China)

Lan LI

Board member

The Hong Kong Polytechnic University

(Hong Kong)

Yukio TONO

Board member

Tokyo University of Foreign Studies

(Japan)

Jirapa VITAYAPIRAK

Convener, ASIALEX2018

King Mongkut's Institute of Technology Ladkrabang

(Thailand)

Mehmet GURLEK

Convener, ASIALEX2019

Istanbul University

(Turkey)

Hai XU

Co-Chief Editor of LEXICOGRAPHY

Guangdong University of Foreign Studies

(China)

Shigeru YAMADA

Co-Chief Editor of LEXICOGRAPHY

Waseda University

(Japan)

Ilan KERNERMAN

Past President

K Dictionaries

(Israel)

Organising Committee

Chair:

Dr Mehmet Grlek

Co-Chair:

Asst Yasin Tademir

Members:

Dr Mehmet Aygne

Dr Murat Elmalı

Dr Ahmet Isparta

Dr Ahmet Naim iekler

Dr Serpil Tuner

Asst Pınar Karakılık

Asst Seza Tabaklar

Asst Abdullah Topraksoy

Scientific Committee

Dr Ezgi ASLAN, Anadolu University, Turkey

Dr Mehmet AYGÜNEŞ Istanbul University, Turkey

Prof Dr Erdoğan BOZ, Eskişehir Osmangazi University, Turkey

Dr Ferdi BOZKURT, Anadolu University, Turkey

Assoc Prof Dr Shirley DITA De La Salle University, The Philippines

Dr Fatih DOĞRU, Eskişehir Osmangazi University, Turkey

Prof Dr Yongwei GAO Fudan University, China

Dr Duygu KAMACI GENCER, Eskişehir Osmangazi University, Turkey

Dr Mehmet GURLEK, Istanbul University, Turkey

Ilan KERNERMAN, K Dictionaries

Dr Deny KWARY, Universitas Airlangga, Indonesia

Assoc Prof Dr Lan LI , The Hong Kong Polytechnic University, Hong Kong

Dr Rachel Edita O. ROXAS, National University, The Philippines

Assoc Prof Dr Vincent OOI , National University of Singapore, Singapore

Assoc Prof Dr Bülent ÖZKAN, Mersin University, Turkey

Prof Dr Yukio TONO, Tokyo University of Foreign Studies, Japan

Assoc Prof Dr Jirapa VITAYAPIRAK, King Mongkut's Institute of Technology Ladkrabang, Thailand

Prof Dr Hai XU, Guangdong University of Foreign Studies, China

Prof Dr Shigeru YAMADA, Waseda University, Japan

Contents

THE TREATMENT OF KOREAN ONOMATOPOEIA-MIMESIS IN KOREAN-INDONESIAN DICTIONARIES <i>Achmad Rio Dessiar Sri Wahyuningsih, Kilim Nam</i>	3
TERMINOLOGY OF ART MUSIC IN OTTOMAN TURKISH AND MODERN TURKISH LEXICOGRAPHY (19TH-21ST CENTURY) <i>Agata Pawlina</i>	22
SWITCHING FROM ARABIC LEXICOGRAPHICAL TRADITION TO RUSSIAN: CASE STUDY – TATAR DICTIONARIES <i>Alina Minsafina</i>	35
TURKISH LEARNER’S DICTIONARY <i>Anna Golynskaia</i>	43
ENRICHING SYNSETS IN TAMIL WORDNET: PARADIGM SHIFTS IN LEXICOGRAPHY <i>S.Arulmozi</i>	54
DICTIONARY OF ANCIENT TURKIC AND MONUMENTS OF MONGOLIAN RUNIC INSCRIPTION <i>Azzaya Badam, Otgonsuren Tseden</i>	63
HISTORY AND DEVELOPMENT OF DICTIONARIES ON INDIGENOUS ENDANGERED LANGUAGES OF CENTRAL INDIA: THEIR PAST, PRESENT AND FUTURE <i>Mendem Bapuji</i>	80
THE METHOD OF THE REAL LIFE BASED SCHOOL DICTIONARY <i>Bülent Özkan, Ferdi Bozkurt, Nurettin Demir, Erdoğan Boz, Şükrü Halûk Akalın</i>	94
BERBER LEXICOGRAPHY: SEMANTIC AND MORPHOLOGICAL PROBLEMS <i>Carla Ferrerós Pagès</i>	106
ON LEXICAL EQUIVALENCE IN THE BILINGUAL DICTIONARY PHILOSOPHICAL AND MENTAL-REPRESENTATION REFLECTIONS: <i>Cuilian Zhao</i>	126
A PERSPECTIVE ON THE PAST, PRESENT AND FUTURE OF LEXICOGRAPHY WITH SPECIFIC REFERENCE TO AFRICA <i>D.J. Prinsloo</i>	148
CONSIDERATIONS FOR PROVIDING ETYMOLOGICAL INFORMATION IN THE KBBI INDONESIAN DICTIONARY <i>David Moeljadi, Ian Kamajaya, Azhari Dasman Darnis</i>	161
SIGNIFICANCE OF PHRASEOLOGY FOR LEXICOGRAPHY AND PHRASE MINING BASED ON CHINESE CORPUS <i>Dejun Li, Man Fu</i>	179
TRILINGUAL ENGLISH-FILIPINO-PAMPANGO GLOSSARY OF CARPENTRY TERMS: BASIS FOR K-12 MATERIALS DESIGN <i>Eliezer V. David</i>	197
DO AFRICAN LANGUAGE CHILDREN’S DICTIONARIES MEET THE NEEDS OF THEIR TARGET USERS? <i>Elsabé Taljard, D. J. Prinsloo</i>	212
INTERACTIVE TERM DEFINING MODULE: A MODAL OF LEXICOGRAPHY TERMS DICTIONARY: <i>Erdoğan Boz, Bülent Özkan, Nilay Girişen</i>	217
GUILLAUME BUDÉ: AN EFFICIENT TRANSFORMER OF GREEK LEXICOGRAPHY IN EARLY MODERN EUROPE <i>Erman Gören</i>	224
THE USER IS KING: ADVICE TO LEXICOGRAPHERS OF LEARNER’S DICTIONARIES <i>Donna M.T.Cr. Farina, Marjeta Vrbinc, Alenka Vrbinc</i>	240

WESTERN LOANWORDS DERIVED WITH TURKISH SUFFIXES IN TURKISH DICTIONARIES <i>Fatih Doğru</i>	249
A BIBLIOMETRIC STUDY OF LEXIKOS <i>Ferdi Bozkurt</i>	270
ON XML-MEDIAWIKI RESOURCES, ENDANGERED LANGUAGES AND TEI COMPATIBILITY, MULTILINGUAL DICTIONARIES FOR ENDANGERED LANGUAGES <i>Jack Rueter, Mika Hämäläinen</i>	284
CORPUS-BASED TERMINOLOGICAL DICTIONARY OF MUSIC: A CASE STUDY OF ROCK GUITAR <i>Nantakarn Impong, Jirapa Vitayapirak</i>	291
COUNT-BASED SEMANTIC MODEL EVALUATION FOR THE EXTRACTION OF SEMANTIC RELATIONS FOR NAMED BAYS FROM A SMALL SPECIALIZED CORPUS <i>Juan Rojas-Garcia, Riza Batista-Navarro</i>	302
DICTIONARY OF OLD RUSSIAN PLANT NAMES (11TH–17TH CC.): WORD ENTRIES DRAFTS <i>Kira I. Kovalenko, Valeria B. Kolosova</i>	316
DICTIONARIES, CORPORA AND ARCHAIC WORDS: THE CHANGE OF CHINESE CHARACTERS WITH THE WOMAN RADICAL <i>Lan Li</i>	329
AUGMENTING CROSS-LINGUAL TERMINOLOGIES WITH TREE-TO-SEQUENCE NEURAL MACHINE TRANSLATION <i>Long-Huei Chen, Kyo Kageura</i>	342
THE ROLE OF CORPORA AND E-LEXICOGRAPHY IN THE DIDACTICS OF FRENCH PHRASEOLOGY <i>Mariangela Albano</i>	352
FINNISH–TURKISH DICTIONARIES: PRESENT STATE AND FUTURE PERSPECTIVES <i>Mats-Peter Sundström</i>	376
ON THE FUTURE OF LEXICOGRAPHY AND DICTIONARIES IN TSHIVENḐA <i>Munzhedzi James Mafela</i>	384
TOWARDS DIGITAL DICTIONARIES HAVING MORPHOLOGICAL ANALYSIS <i>Mutee U Rahman, Tafseer Ahmed</i>	391
CONNOTATION VERSUS DENOTATION: THE EFFECTS OF CONNOTATION IN THE LEMMATIZATION OF TERMINOLOGY <i>Dr. MV Mojela</i>	400
BILINGUAL DICTIONARY FOR ACADEMIC WRITING: HEDGES, BOOSTERS AND ATTITUDE MARKERS AS INTERACTIONAL RESOURCES <i>Neslihan Onder Ozdemir</i>	408
TRANSFORMING GLOSSARIES INTO KNOWLEDGE RESOURCES: FRAME-BASED TERMINOLOGY APPLIED TO MILITARY SCIENCE <i>Pamela Faber, Pilar León-Araúz</i>	412
THE NAME ‘NDEBELE’ CAN SUGGEST THE SAME OR A DIFFERENT LINGUISTIC GROUP <i>Dr. K.S Mahlangu</i>	439
TEXT DATA INDEXING IN LEXICOGRAPHIC STUDIES: AN APACHE SOLR APPLICATION <i>B. Tahir Tahiroğlu</i>	450
AN OVERVIEW ON THE HISTORY OF RUSSIAN LEXICOGRAPHY <i>Ümmügülsüm Dohman</i>	457
CREATING A TRILINGUAL DICTIONARY FOR WESTERN YUGUR, AN ENDANGERED TURKIC LANGUAGE <i>Yarjis Xueqing Zhong</i>	464

A CRITICAL REVIEW OF THE INCLUSION OF ACRONYMS AND INITIALISMS IN <i>THE ENGLISH-CHINESE DICTIONARY</i> <i>Yongwei Gao</i>	473
HOW DO PEOPLE UNDERSTAND THE 'AUTHORITY' OF AN ENGLISH DICTIONARY? ANALYSIS OF PEOPLE'S REACTIONS TO THE INCLUSION OF NEW WORDS <i>Yuri Komuro</i>	485
A STUDY OF THE REPRESENTATION OF VERB-NOUN HETEROSEMY IN <i>THE CONTEMPORARY CHINESE DICTIONARY (7TH ED.)</i> <i>Yushuang Dong, Renqiang Wang</i>	491
THE TREATMENT OF PHRASEOLOGY IN CHINESE-ENGLISH DICTIONARIES: A PRELIMINARY STUDY <i>Zhang Xuhua</i>	514
LMF RELOADED	
<i>Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, Piotr Bański</i>	533
FROM CONCEPT DEFINITIONS TO SEMANTIC ROLE LABELING IN SPECIALIZED KNOWLEDGE RESOURCES <i>Ivana Brač, Ana Ostroški Anić</i>	540

PREFACE

The 13th International Conference of the Asian Association for Lexicography (ASIALEX 2019) is hosted by İstanbul University. The theme of the conference is Lexicography: Past, Present, Future. We are happy to have four distinguished keynote speakers: Amy Chi from Hong Kong University of Science and Technology, Danie Prinsloo from the University of Pretoria, Gilles-Maurice de Schryver from Ghent University and Hayati Develi from İstanbul University.

After passing through a double-blind review by the Scientific Committee, 85 papers were accepted for presentation at the AsiaLex2019. The Conference Proceedings have compiled a total of 42 papers

It is hoped that this collection of papers presented at the AsiaLex2019 serve as a useful resource for future researchers on lexicography and other related fields.

Dr. Mehmet Gürlek

Academic Convenour, ASIALEX 2019

THE TREATMENT OF KOREAN ONOMATOPOEIA-MIMESIS IN KOREAN-INDONESIAN DICTIONARIES

Achmad Rio Dessiar

Sri Wahyuningsih

Kilim Nam

Kyungpook National University

Abstract

This study identifies issues regarding the treatment of Korean onomatopes and mimetic words in Korean-Indonesian dictionaries (printed and online). The Korean language is characterized by a rich sound-symbolic system, which is reflected in Korean monolingual dictionaries with the inclusion of many onomatopoeic and mimetic headwords. However, the description of sound-symbolic forms is relatively poor in Korean-Indonesian bilingual dictionaries in terms of both quantity and quality. In order to determine the reasons for such shortcomings, it has been necessary to analyze the Korean sound-symbolic word entries. We collected the data from *Korean Onomatopoeia and Mimesis* (Park, 2012), which presents 202 onomatopoeic and mimetic words. Then, we looked up their translations in the printed *Korean-Indonesian Dictionary* (Jung, 1998) and on Naver online Korean-Indonesian dictionary.

We identified four types of issues regarding the treatment of onomatopes and mimetic words in the above dictionaries and classified them as follows: 1) POS shifting; 2) inaccurate translation; 3) zero translation; 4) lack of systematicity. We could observe that shortcomings were more common in the printed dictionary than the online version. POS shifting occurred in 64 out of 202 lemmas. In the case of the printed dictionary, there were 31 cases where the onomatope or mimetic word was only presented in derivative forms. Finally, inaccurate translations could be found across 10 entries in both printed and online dictionaries. In Naver online dictionary in particular, the descriptions of onomatopoeic and mimetic headwords tend to be simple translations of the corresponding definitions provided by the *Standard Korean Language Dictionary*. We hope that the results of this study will not only contribute to improving the Korean-Indonesian and Indonesian-Korean lexicography, but also spark interest and lead to further research in cross-linguistic onomatopoeia and mimetic words.

Key Words: Dictionary Treatment; Korean Onomatopoeia; Mimesis; Bilingual Dictionary; Translation

I. INTRODUCTION

The purpose of this research is to identify the treatment regarding the translation of Korean onomatopoeic-mimetic words (OM)¹ and its related issues in the Korean-Indonesian dictionaries, both printed and online dictionaries. There are many difficulties experienced by foreigners when learning Korean. One reason for difficulties is the large number of onomatopoeic and mimetic expressions. Onomatopoeia can be defined as words imitating the natural sounds. Meanwhile, mimetic words are word mimicking the shapes or state of objects (Yoon, 1993:14). In Korean language OM is categorized as sound symbolic words. The Korean language is characterized by a rich sound-symbolic system, which is reflected in Korean monolingual dictionaries with the inclusion of many onomatopoeic and mimetic headwords

Every language throughout the world possesses OM. A language is initially constructed by the formation of OM, which is a symbol of perception and humans' basic behavior (Chae, 2003: 5). OM in each language generally has a variety of characteristics. In Korean, OM are particularly common in the form of repetition. OM can be classified based on their compound into three types: single word forms, compound word forms, and derivative forms. OM are basically categorized as adverbs because they tend to modify a predicate in a sentence. Furthermore, OM in Korean, which is attached with suffixes, causes word derivation process. The derivational process then generates a variety of parts-of-speech: derivational noun, derivational verb, and derivational adjective. The wide variety of OM forms undoubtedly cause difficulties in determining the most appropriate headword to be included in dictionaries. It will also be increasingly difficult to translate the OM forms into the target language.

OM in Korean belongs to adverb that has a fairly high frequency of usage in both formal and informal language². In fact, formal language use of the OM is reflected in questions in the Test of Proficiency in Korean (TOPIK). There are always questions that are related to Korean OM. Therefore, OM are important materials for Korean foreign learners. However, OM materials are often excluded in Korean language teaching for foreign learners. This situation is proven by the limited number of books that explain the OM in detail. It seems that OM is considered less important because it belongs to adverb. In this case, its function in a sentence does not have much influence. Either OM exists or not, this will not significantly change the meaning of the sentence.

When Korean language learners encounter difficulties in understanding OM, dictionaries are very helpful. Proper dictionary treatment definitely needed in order to understand Korean OM. Through this study, we are trying to identify the treatment of Korean OM and other related issues on Korean-Indonesian dictionaries. The research findings are expected to be used in the development of Korean-Indonesian dictionaries in the future. In addition, this research will then be used to formulate teaching materials.

This study is presented in five chapters. The first chapter contains an introduction, followed by the second chapter which consists of research methods. In the third chapter, the result of the study will be presented. Chapter 4 discusses the translation of Korean OM in Korean-Indonesian dictionaries, both printed and online. Chapter 5 contains conclusion and suggestions related to the result.

¹Henceforth the word onomatopoeic-mimetic words will be abbreviated as OM.

²The number of Korean onomatopoeic-mimetic word entries in the dictionary is 8,283, compared to other languages, Korea has a large amount of onomatopoeia (Park, 2012: 4).

II. METHODS

This research was carried out in several stages. The first step was data collection. The research data are OM in Korean that are often used in daily life. The amount of data used is 202 OM which are all taken from the *Korean Onomatopoeia and Mimesis* book (Park, 2012). This book can be considered as an OM dictionary created to support Korean language teaching for foreigners and children from mixed marriages in Korea.

The next stage of this research was data analysis. Among 202 forms of OM, researchers identified their definitions and usage in Standard Korean Language Dictionary. Afterward, the researcher identifies their Indonesian translations in Korean–Indonesian dictionaries both printed and online. In this research, the Standard Korean Language Dictionary used is the online one (<https://stdict.korean.go.kr/main/main.do>). Meanwhile, the Korean–Indonesian dictionaries used are the printed dictionary compiled by Hankuk University of Foreign Studies (HUFs, 1998) and the online Naver dictionary (<https://dict.naver.com/idkodict/#/main>). Both of them are chosen because they are the earliest developed dictionaries and have most word entries with the most complete word descriptions, compared to other Korean–Indonesian dictionaries.

The results of identified OM translation were registered into Microsoft Excel to be analyzed. The researchers then identified the issues that arise in the treatment of Korean OM in Korean–Indonesian dictionaries. The final stage of this research is classifying issues regarding Korean OM translation and headword determination in the Korean–Indonesian dictionaries.

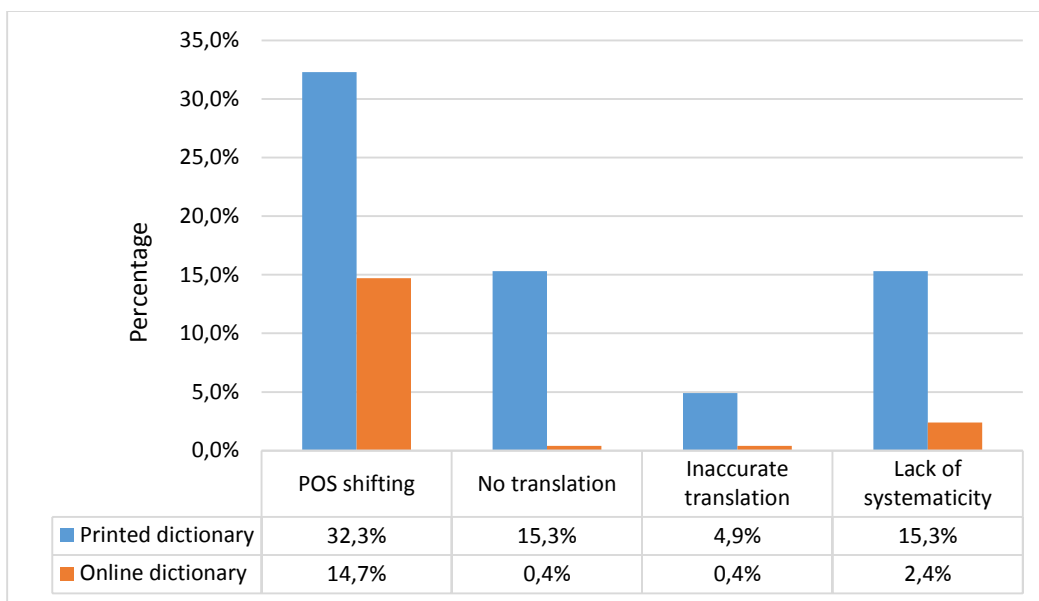
III. RESULTS

Based on the analysis of OM translations in both printed and online Korean–Indonesian dictionaries, OM treatment can broadly be classified into four categories as shown in Table 1 and Graph 1 below.

Table 1. Analysis Result of Korean OM in Korean-Indonesian Dictionaries

	POS shifting (per case)	Inaccurate translation	No translation	Lack of systematicity
Printed dictionary	66	10	31	31
Online dictionary	30	1	2	5

Graph 1. Analysis Result of Korean OM in Korean-Indonesian Dictionaries



IV. DISCUSSION

This section will discuss the issues regarding translation and headword determination of OM both in printed and online dictionaries. Mostly, the issues discussed are classified into four categories: part-of-speech (POS) shifting, inaccurate translation, having no translation (zero translation), and lack of systematicity.

IV.1. Part-of-Speech (POS) Shifting

OM is basically an adverb which function is to modify predicate in a sentence. In Korean, an OM attached to affixes $\sim i^3$ and $\sim bo$, $\sim ami$, $\sim kuri$ forms a noun (nominalization); to affixes $\sim hada$; $\sim georida$, $\sim daeda$, and $\sim ida$ forms a verb (verbalization); and to affixes $\sim hada$, and $\sim reobta$ forms an adjective (Chae 2003: 50-52). Examples of those derivational forms can be seen in Table 2 below. The wide variety of OM forms undoubtedly cause difficulties in determining the most appropriate word entries to be included in dictionaries. Furthermore, it will also be increasingly difficult to translate OM forms into the target language.

Table 2. Derivational Form of Korean OM

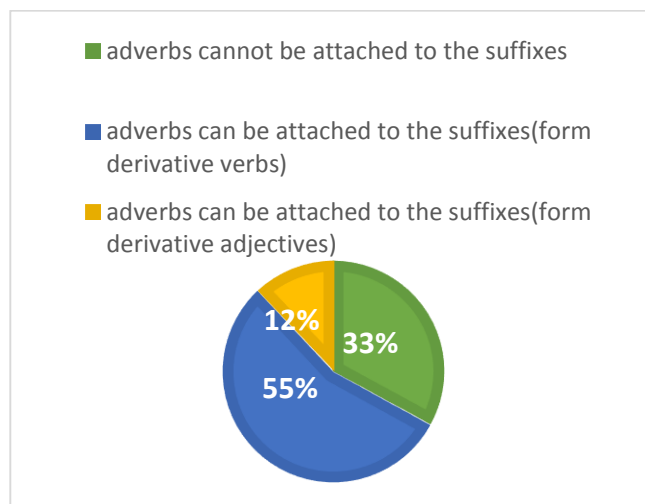
No	Root	Root + Suffix	Type of Derivational
1	<i>meong-meong</i> (a word imitating the sound of a dog barking)	<i>meong-meong <u>\sim i</u></i> (dog)	derivational noun
2	<i>gwittur</i>	<i>gwittur <u>\sim ami</u></i> (cricket)	

³ Korean Romanization system used in this paper is Revised Romanization of Korean

	(a word imitating the sound made by a cricket)		
3	<i>khong-khong</i> (a word imitating the sound made when a small, light object falls on the ground or on another object.)	<i>khong-khong~georida</i> (thump)	derivational verb
4	<i>kkal-kkal</i> (a word imitating the sound of laughing in a loud voice.)	<i>kkal-kkal~daeda</i> (guffaw)	
5	<i>budeur</i> (in a state in which a feeling to the skin is very smooth/ softly.)	<i>budeur~reobda</i> (soft; smooth)	derivational adjective
6	<i>donggeul-donggeul</i> (in a state in which all of several things are very round/ roundly)	<i>donggeul-donggeul~hada</i> (round)	

Of the 202 data entries used in this research, 67 (33%) belong to adverbs and cannot be attached to suffixes. The remaining 135 data entries (67%) belong to adverb that can be attached to suffixes and form derivative words. Among those 135 entries, 111 words generate derivative verbs, while the other 24 words form derivative adjectives as shown in Chart 1.

Chart 1. Korean OM Form



In Korean dictionaries, the OM head words are commonly found as the root words⁴. The root of the OM is an adverb that modifies the predicate in a sentence. POS shifting is the most common finding in the OM translations the Korean–Indonesian dictionaries. Based on the results of the analysis, each printed and online dictionary recorded 32.3% and 14.7 % cases that were, not translated into adverbs. This clearly proves that parts-of-speech shifting is the most common finding in this research. The translation results that encounter POS shifting can be seen in Table 3 below.

Table 3. Part-of-Speech (POS) Shifting found in Printed and Online Dictionaries

	Verb	Adjective	Noun	Total
Printed dictionary	55%	30%	15%	100%
Online dictionary	43%	47%	10%	100%

From the total number of POS shifting cases in printed dictionary, the shift to verbs is 55%, to adjectives is of 30%, and to nouns is 15%. Furthermore, in the online dictionary, the derivation to verbs is 43%, to adjectives is 47%, and to nouns is 10%.

IV.1.1. POS Shifting into Verbs

The shift of adverbs to verbs occupies the top position in OM translations in the printed dictionary and occupies the second position in the translations of online dictionary. This generally occurs when the translation involves OM that can be attached to suffixes and form verbalization. An interesting finding in this case is that the Korean OMs’ translation into Indonesian verbs does not only involve those OM that can be attached to the suffixes. Hence, it also involves the OM that basically cannot be attached to the suffixes but are very closely associated with particular verbs. Therefore, in this case, the translation of the OMs tends to be the translation of the verb itself.

For example, in the printed dictionary, the word ‘*bing*’ is a mimetic word expressing a circular shape or state that always attached to certain verbs such as ‘*dolda*’ (*id: berkeliling*; *en: to get around*).⁵ The printed Korean-Indonesian dictionary does not translate the word ‘*bing*’, but instead translates the verb ‘*dolda*’ which means to get around (*berkeliling*). Also, in the online dictionary, the word ‘*ssik*’ which is a mimetic word mimicing the state of smile that always attached to the verb ‘*utta*’ (*id: tertawa*; *en: to laugh*), translated as ‘*tertawa*’ (to laugh) which belongs to the verbs in Indonesian.

**Figure 1. Examples of the OMs Translation into Verbs
in Korean–Indonesian Dictionaries⁶**

⁴ Based on the Standard Korean Language Dictionary, root word is identified as the central part that represents the actual meaning of the words. A root is a word that does not have a prefix in front of the word or a suffix at the end of the word (Kemmer, Suzzane). The root word primary lexical unit of a word, ad of a word family (this root then called the base word), which carries the most significant aspects of semantic content and cannot be reduced into smaller constituents.

⁵ *id*: Indonesian; *en*: English

⁶ In this paper, examples of entries in dictionary, both printed and online dictionaries are presented in Figure 1~9. For boldly italic letter are romanization of onomatopoeic and mimetic word in

Printed dictionary

bing ① [*han bakki doneun moyang*] *berputar* (to spin around); *berkeliling* (to get around).

¶ [*Seoule han bakhi bing dolda berkeliling Seoul*] (to get around Seoul) ② [*eulossageona*

anjta] *melingkar* (to make a circle). ¶ [*ping dulleo anjda duduk melingkar*] (to sit in a circle). ③ [*Jeongsini*] ¶ [*meoriga ping dolda merasa pusing*] (to feel dizzy)

④ [*neunmuri*] ¶ [*nunmuri ping dolda merasa ingin menangis*] (want to cry).

Online dictionary

*ssik*1. *tertawa satu kali dengan menaikkan ujung mulut tanpa suara*

(to smile)

4.1.2 POS Shifting into Adjectives

Furthermore, the shift of adverbs into adjectives accounts for most of the issues in the online dictionary and the second most issues in the printed one. Similar to the previous case, this kind of shift (into adjectives) generally occurs when translating the Korean OM that can be attached to the suffixes and form derivational adjectives. Besides, it is found that some forms of adverbs that cannot be attached to suffixes, but closely followed by certain adjectives to be translated to adjectives, as shown in Figure 2.

Korean, italicized words are the translation of Korean onomatopoeic and mimetic words, or other expression in Indonesian.

**Figure 2. Examples of the OMs Translation into Adjectives
in Korean–Indonesian Dictionaries**

<p>Printed dictionary</p> <p><i>teong</i> ¶ <i>Teong bin kosong</i> (blank, empty); <i>hampa</i> (blank, empty); <i>melompong</i> (blank, empty)/ <i>teong bin neukim rasa kekosongan</i> (solitude); <i>perasaan kosong (dalam diri)</i> (loneliness); <i>kekosongan</i> (emptiness)/ <u><i>bangi teong bida kamar kosong</i></u> (an empty room)/ <i>Hojumeoniga teong bida dompet tidak berisi sepeserpun</i> (an empty wallet)</p>
<p>Online dictionary</p> <p><i>hwaljjak</i> 1. <i>lebar</i> (wide)</p> <p>2. <i>bentuk kondisi membentangkan lebar sayap atau lengan</i> (to widely spread a wing or a hand)</p> <p>3. <i>bentuk kondisi sesuatu terbentang dengan luas</i> (being widely spreaded out)</p>

As the examples shown above, the word ‘*teong*’ stated in the printed dictionary is a mimetic word for expressing an empty situation and always attached to the adjective ‘*bida*’, which means ‘empty’ (*kosong*). The Korean–Indonesian printed dictionary does not directly translate the word ‘*teong*’, but instead, translates the adjective ‘*bida*’ which means ‘empty’ (*kosong*). Furthermore, the example presented in the online dictionary shows that the word ‘*hwaljjak*’, which is a mimetic word for expressing a wide-spread situation, is translated in Indonesian as the word ‘*lebar*’ (wide), which is an adjective.

4.1.3. POS Shifting into Nouns

The OM translation resulting in a POS shifting of adverbs to nouns shows the fewest cases, both in printed and online dictionaries. The form shifting from adverbs into nouns can be seen in the example in Figure 3.

**Figure 3. Examples of OMs Translation into Nouns
in Korean–Indonesian Dictionaries**

<p>Printed Dictionary</p> <p><i>peong-peong</i> ① 【Pugeum sori】 <i>letusan</i> (eruption, explosion); <i>ledakan</i> (explosion, bang); <i>letupan</i> (eruption, explosion). ② 【ssodajineun moyang】 <i>pancaran</i> (gush, spurt); 【nuni】 <i>dalam pusingan penyemburan; pencurahan</i> (spurt, gush, squirt). ¶ <i>~sodneunsem air terjun yang memancar</i> (a pouring waterfall)/ <i>~heureuda memancar/ nuni ~ssodajida salju turun dengan lebat dan cepat</i> (snow pour heavily and quickly)/ <i>nunmureul ~ ssodda mengalir air mata</i> (streaming tears); <i>banyak sekali mencururkan air mata</i> (tears stream a lot)/ <i>piga sangcheo-eseo ~ssodajyeo nawatta Darah memancar dr lukany</i> (Blood gushes from the wound).</p>
<p>Online Dictionary</p> <p><i>oksin-kaksin</i> 1. <i>percekcokan</i> (a dispute, a conflict)</p> <p>1. <i>bertikai secara lisan sambil saling berdebat siapa yang salah siapa yang benar atau dengan mengemukakan pendapat sendiri</i> (to quarrel, to argue, to fight)</p>

The word ‘*peong-peong*’ is an onomatopoeic word that imitates bursting or emitting sounds. In the printed dictionary, the word ‘*peong-peong*’ is translated as ‘*letupan*’ (a burst) and ‘*pancaran*’ (a gush), which in Indonesian is a noun. Then, there is the word ‘*oksin-kaksin*’ which is a mimetic word that expresses the situation of fights and always uses with verbs related to fights such as ‘*dattuda*’ (to quarrel), ‘*Ssauda*’ (to fight/ to fist), or ‘*maldattumhada*’ (to argue). In this translation, the online Korean–Indonesian dictionary does not directly translate the word ‘*oksin-kaksin*’ as an adverb, but instead translates it with the noun ‘*percekcokan*’ (a dispute, a conflict).

This study also found that one head word undergoes more than one case of POS shifting both in the printed and online dictionaries. This implies that in a set of translations of an OM, it is possible to produce several words possessing varied POSs. To clarify, examples are presented in Figure 4.

Figure 4. Examples of OMs Translation Encountering More Than One Form Shifting in Korean–Indonesian Dictionaries

Printed Dictionary

kam-kam *ketidaktahuan* (ignorance); *kebodohan* (foolishness); *kedunguan* (stupidity); *kosong* (blank; empty); *hampa* (blank; empty); *blanko* (a formulir). **-hada** *tidak mengetahui* (neglectful); *bodoh* (foolish); *dungu* (stupid); *bebal* (ignorant); **【eodubta】** *sangat gelap* (completely dark). ¶**~han eodum** *kegelapan sekali* (an absolute darkness); *sangat gelap* (completely dark)/ **~han sogeseo** *dalam kegelapan yg sama sekali* (in a complete darkness)/ **-ida** *tidak mengetahui tentang ((hukum))* (worst at law)/ **naneun hanhakeneun jeonhyeo ~ida** *Saya sama sekali tidak mengetahui tentang sastra klasik Cina* (I have no idea at Chinese classic literatures). / **geureon irieneun aju(jeonhyeo) ~ida** *Saya tidak mengetahui apa-apa mengenai persoalan-persoalan sama itu* (I know nothing at all about the problems with her/him)

Online Dictionary

tabak-tabak 1. *Bentuk berjalan lambat tanpa tenaga* (a situation of walking slowly and powerlessly)

tabak-tabak² 1. *Kering* (dry) 2. *Mengering* (to dry up) 3. *Tidak lembab* (not moist)

The table above shows that the word 'kam-kam' is translated into 'ketidaktahuan' (ignorance), 'kebodohan' (foolishness), and 'kedunguan' (stupidity), each of which belongs to the noun. Furthermore, it is also translated into 'kosong' (blank) and 'hampa' (empty), which are adjectives. In the online dictionary, the word 'tabak-tabak' on the one hand, translated into Indonesian becomes 'kering' (dry) and 'tidak lembab' (not moist), which are adjective, and becomes the word 'mengering' (to dry up) which is a verb.

IV.2. Inaccurate Translation

A translation can be considered inaccurate if it has different meanings with Standard Korean Language Dictionary. This research found the inaccuracy in the OMs translation is up to 4.9% in the printed dictionary and 0.4% in the online one.

Figure 5. Examples of Inaccurate Translations in Korean–Indonesian Dictionaries

<p>Printed Dictionary</p> <p><i>ttuk</i> ① 【tteoreojineun moyang】 <i>bergedebug</i> (thud) ② 【bureojigeona kkeuneojineun moyang】 <i>berderik</i> (rattling). ¶<i>hobaki ttange ttuk tteoreojida</i> sebuah labu jatuh dengan bunyi gedebug di atas tanah (a pumpkin falls thud on the ground)/ <i>ttuk bureojida pecah</i> (patah) dengan bunyi yang keras (to be broken with a loud noise)/ <i>sireun ttuk kkeunta</i> memutuskan seutas benang menjadi dua bagian (to cut off a piece of yarn into two pieces)/ <i>daedeulboga ttuk haetta</i> Tiang besar (utama) itu berbunyi keriat-keriut (The big main pole trills)/ <i>batjureui gajang yakhan gosi ttuk kkeuneojyeotta</i> Tali itu putus pada ujung yang paling lemah (The rope is cut off at its weakest end.) ...</p> <p><i>kkobak-kkobak</i> ① <i>kkubo~kkubok</i> ② 【sunjung】 tanpa kelalaian (accurately); tanpa kegagalan (successfully) ¶<i>segeumeul ~ chireuda</i> membayar pajak secara teratur (to pay taxes regularly) ③ 【gidarim】 menunggu dengan hati bimbang (to wait for with a wavering heart).</p>
<p>Online Dictionary</p> <p><i>hwak</i> 1.kondisi angin (condition of wind) 2.bau (a smell) 3.semangat (a spirit)</p>

In this study, a translation is said to be inaccurate if it contains at least one ambiguous word in the target language. It can be seen in Figure 5 that the word ‘*ttuk*’ and the word ‘*berderik*’ (rattling) in Indonesian is not equal. According to The Great Dictionary of Indonesian Language (id. KBBI), ‘*berderik*’ (rattling) has exactly the meaning of ‘*tiruan bunyi papan bergesekan*’ (an imitating sound of board friction). On the contrary, it can be ensured that in the Standard Korean Language Dictionary, there is not even a single definition of ‘*ttuk*’ appropriate to the Indonesian word ‘*berderik*’ (rattling). Most of the definitions of ‘*ttuk*’ explain the situation of an object that either falls or stops suddenly. Another example is the word ‘*kkobak-kkobak*’ which translates to ‘*menunggu dengan hati bimbang*’ (to wait for with a wavering heart). On the other hand, ‘*kkobak-kkobak*’, according to the Standard Korean Language Dictionary does not even mean ‘*menunggu dengan hati bimbang*’ (to wait for with a wavering heart) at all.

Researchers found ten cases of inaccurate OM translation in the printed dictionary. Whereas, in online Naver dictionary, only one inaccurate OM translation was found, which is the word ‘*hwak*’. This word is mimesis which expresses the condition of wind or smell that quickly or suddenly arises. In the online dictionary, ‘*hwak*’ is translated into ‘*kondisi angin*’ (condition of wind), ‘*bau*’ (a smell), and ‘*semangat*’ (a spirit). The use of ‘*hwak*’ as an adverb in Korean that surely can modify the predicate is irrelevant to the

words ‘*kondisi angin*’ (condition of wind), ‘*bau*’ (a smell), and ‘*semangat*’ (a spirit) which obviously cannot modify the predicate in a sentence. This translation, therefore, will clearly confuse the dictionary users.

Table 4. Inacurate OM Translation on Korean-Indonesian Dictionaries

No.	OM	Definition ⁷	Inacurate translation
1.	<i>kkeombeok</i>	1. A word describing the manner of a strong source of light suddenly dimming and then becoming bright. 2. A word describing the motion of a person’s big eyes being close and then open	<i>mati</i> (turned off)
2.	<i>kkobak- kkobak</i>	1. niddle-noddle or bowing repeatedly. 2. In the manner of continuously doing something without skipping it even once	<i>menunggu dengan hati bimbang</i> (to wait for with a wavering heart)
3.	<i>nabjak- nabjak</i>	1. with one’s mouth open. 2. In the manner of flattening one’s body to a floor or wall	<i>dengan gerakan cepat</i> (with a very quick movement)
4.	<i>deobseok</i>	A word describing a person or an animal suddenly darting at something and biting or grabbing it.	<i>menghardik</i> (scold)
5.	<i>ttuk</i>	1. A word imitating the sound or describing the motion of big object or waterdrop, etc falling. 2. A word imitating the sound or describing the motion of big, hard object being broken or cut. 3. A word imitating the sound or describing the motion of hitting a hard object one time. 4. In the manner of picking or nipping something off something else it is clinging or attached to without hesitation.	<i>berderik</i> (ratling)
6.	<i>basak</i>	1. A word imitating the sound made when a dry leaf or tree bough, etc, is stepped on lightly or describing that scene. 2. A word imitating the sound made when a dry object touches each other or are chrused. 3. A word imitating the sound made when a small and hard object is crushed or broken, or describing that scene	<i>gercik</i> (splatter)
7.	<i>Umul-umul</i>	1. A word describing someone talking indistinctly as if muttering. 2. In the state of not swallowing but instead continuing to chew food. 3. A word describing someone’s lip, muscle, etc, contracting	<i>berkerumun, mengerumuni</i> (crowd around)

⁷ Definition based on Standard Korean Language Dictionary

		repeatedly. 4. A word describing someone acting in an uncertain or hesistant manner	
8.	<i>umpuk</i>	1. In the stat of being hollowed out or curving inwards in the middle	<i>dalam lembah (lubang)</i> (inside the valley (hole))
9.	<i>hukkeun</i>	1. In the state of an object becoming suddenly hot, being exposed to heat or one's face becoming suddenly hot and turning red. 2. The state of excitement or tension, etc., becoming stronger	<i>merasa puas</i> (satisfied)
10.	<i>heulkit-heulkit</i>	A word describing the motion of looking askance at something quickly one time	<i>membelalang dan membelalang, terus-menerus memandang dengan marah</i> (stare and glare, constantly looking angrily)
11.	<i>hwak</i>	1. In the manner of wind, a smell, or energy emerging suddenly and strongly. 2. A word describing th emotion of fire surging sudenly, strongly. 3. In the state of one's face suddenly feeling hot and turning red. 4. In the state of one's face suddenly feeling hot and turning red.	<i>kondisi angin</i> (condition of wind)

Incorrect translation as mentioned above might cause foreign learners and/or dictionary users to misunderstand the meaning of OM. Those inaccurrate translations should be corrected by finding the equal words in Indonesian. But if it is difficult to find the equal words, those OM should be defined according to the Standard Korean Language Dictionary.

4.3. Having No Translation (Zero Translation)

In addition to the two issues described previously: POS shifting; and inaccurate translation, there is one issue left related to OM translation, namely zero translation. As many as 31 OM, or about 15.3% of all the research data, are not found at all in the printed dictionary as headword. Those intended OM can be seen in Table 5 below. Meanwhile, it is found that about 0.4% (2 datas) OM data entries do not have translation on online dictionary, that is word '*gangjung-gangjung* and *jam-jam*'

Table 5. OM Zero Translation on Printed Korean-Indonesian Dictionary

gyauttung; gyaut; geongjung-geongjung; geudeughada; kkangcungllangcung; kkomkkomhi; kkumteul; kketteok; neobjuk; dalgeurak; dung-dung; ttarreureung; bajjak; beolleong-beollong; beolkeot; binggeut; paek-paek; peokkuk; songgol-songgol; sswa; ssik-ssik; jjeut-jjeut; chalsak-chalsak; cheolleong; puljjak-puljjak; heok-heok; hwadeuljjak; hudeudeug; heuikkeun-heuikeun;hilkkeum; hilkkeut

Generally, in the process of compiling the dictionary there are several stages including data collection process. Standards for determining headword and data sources in each dictionary can vary. The absence of OM mentioned above as headwords in Korean-Indonesian dictionaries are most likely because during the process of collection data, those OM words are not selected.

4.4. Lack of Systematicity

In the treatment of Korean OM in Korean-Indonesian Dictionary, it is found that systematicity is lacking in putting headword. Inconsistencies are found in the putting headword of Korean OM as much as 15.3% in the printed dictionary, and 2.4% in the online dictionary. The OM's headword put in the dictionary should have been a root word. Derivational form of root word is ideally inserted as word entry (not as headword) as in Figure 4 above. Figure 4 shows that word 'kkam-kkam' which is headword is a root word. The derivational forms of 'kkam-kkam' ('kkam-kkam+hada; kkam-kkam-ida') are inserted as word entries.

However, in this case, it also implies that the head words are derivative words. For example, the word 'banjjak' attached to the suffix *-ida* (printed dictionary) and the word 'eoseuleong' attached to the suffix '*-daeda*' (online dictionary) are included as the head word. Examples are presented in figure 6 below.

Figure 6. Examples of Lack of Systematicity

<p>Printed Dictionary</p> <p><i>banjjak-ida</i> berkilau (shining), bergemerlapan (sparkling), bercahaya (glowing), bersinar-sinar (twinkling), berbinar (shining). ¶<i>haetbiche geoureul-menyorotkan cermin pada sinar matahari</i> (reflecting the sunshine to the mirror)/ <i>geunyeoeui boseoki nampho bulbichi banjjakyeotta Permatanya bersinar (bercahaya) dalam cahaya lampu</i> (Her diamond is shining under the lights. / <i>meondeseo buri banjjakinda Cahaya yang samar-samar (yang redup) dari kejauhan</i> (the dim lights from distance)/ <i>Puribbe iseuri maecheo(seo) banjjalida Tetsan embun berkilau-kilauan di atas dedaunan</i> (The dew drops are twinkling on the leaves).</p>
<p>Online Dictionary</p> <p><i>eoseuleong~daeda</i> 1. jalan-jalan (to walk around) 2. mondar-mandir (back and forth) 3. berkeliling (to go around)</p>

The inconsistencies described above are most commonly found in the printed dictionary. In the printed dictionary, there are 31 head words inserted in the form of derivation. Five derivational headwords are also found in the online dictionary. When the OM word entries are inserted in the form of derivative words, it can almost be ensured that the POS of the translation result will remain as derivative words. Examples are the words 'banjjak' and 'eoseuleong', which are adverb. After being attached to the suffixes, become verbs,

namely *banjjak-ida* and *eoseuleong-daeda* in the dictionaries, automatically, the translation results of those words in Indonesian will follow and turn into verbs, '*berkilau*' (shining) and '*jalan-jalan*' (to walk around), respectively.

We also found that the root words and all of their derivative words inserted as headwords in dictionary. This is commonly found in the printed dictionary. For example, the word *bulssuk* is not only entered as a root, its derivative forms, which are *bulssuk-hada* and *bulssuk-georida*, are also included as the head word. This process, besides being considered inconsistent, also creates inefficiency.

Figure 7. The Inconsistency of the Head Word in Printed Dictionary

bulssuk dengan tiba-tiba (suddenly, spontaneously); *tiba-tiba* saja (suddenly, spontaneously); *tahu-tahu* (suddenly, spontaneously). ¶~*nathanada* tampil dengan tiba-tiba (suddenly appears)/~*deuleuda* melakukan kunjungan kejutan ((ke)) (to carry out a spontaneous visit to-)/*jumeokeul* ~*naemilda* meninju dengan tiba-tiba (to punch spontaneously)/*cheongbakeuro* ~*naemilda* mengeluarkan kepala ke jendela dengan tiba-tiba (to pull out head through the window suddenly) *eodumsokeseo saram geurinja*~*nathanada* Sesosok tubuh manusia tampak besar di kegelapan (A human body looks giant in the dark).

bulssuk~*georida* ① **[naemilda]** mengeluarkan dengan tiba-tiba (to pull out suddenly). ¶*jumekeul* ~ tetap meninju dengan tiba-tiba (to punch spontaneously). ② **[nohada]** mudah membangkitkan amarahnya (to make angry suddenly); menjadi marah dengan segera (to get angry suddenly)

Bulssuk~*hada* menonjok (to punch); mengembang (to swell)

Apart from inconsistency in putting headword, translation of Korean OM are systematically presented in Korean–Indonesian dictionary. Despite the fact that most Korean OM do not always have their equivalences in Indonesian, some do. In printed dictionary even though they have their equivalent words in Indonesian, those equivalent words are not inserted. This because there are many word translation on Korean-Indonesian printed dictionary taken from Korean-English dictionary (Chun, 2003: 165-169). For example, in Figure 8, the word '*keol-keol*' is an onomatopoeia that imitates the sound of laughter and '*meong-meong*' is an onomatopoeia that imitates the sound of a dog barking. Those two onomatopoeias have their equivalent in Indonesian. However, in the printed dictionaries, they are not translated in their Indonesian equivalent onomatopoeias.

In the online Korean–Indonesian dictionary, the translation method used is both translating directly into Indonesian onomatopoeias and translating Korean OM definition based on the Standard Korean Language Dictionary. Examples in Figure 8 present the words '*keol-keol*' and '*meong meong*' which are translated directly into Indonesian onomatopoeias as '*hahaha*' (hahaha) and '*guguk*' (woof-woof), and Figure 9 present words '*bajjak*, *uttuk*, *juruk-juruk*' are translated from Standard Korean Language Dictionary definition.

Figure 8. Examples of Inconsistencies in the OMs Translations

in Korean–Indonesian Dictionaries

Printed Dictionary

keol-keol [*~utta*] *tertawa dengan keras* (to laugh out loud); *tertawa terbahak-bahak* (to laugh out loud).

meong-meong *gonggongan (galak) anjing* (sounds of dog barking). [*~georida*] *menyalak* (to bark); *menggonggong* (to bark). ¶ [*gaega ~jitta*] *seekor anjing menggonggong* (A dog barks).

Online Dictionary

keol-keol 1. *Hahaha* (hahaha)

meong-meong 1. *guk guk* (woof-woof)

Figure 9. Comparison of OM Definition in Korean-Indonesian Online and Standard Korean Language Dictionary

Online Dictionary	Standard Korean Language Dictionary
<p><i>bajjak</i> <i>1. bentuk kering atau menyusut sehingga tidak lagi berair (shape of s.o is drying or shrinking so it is no longer contain any water)</i></p>	<p><i>bajjak</i> <i>mulgiga maeu mareugeona jorabutgeona ta beorineun moyang (shape of s.o is drying or shrinking so it is no longer contain any water)</i></p>
<p><i>uttuk</i> <i>bentuk menjulang tinggi yang sangat tajam (very sharp towering shape)</i></p>	<p><i>uttuk</i> <i>dudeureojige nopi sosa inneun moyang (very sharp towering shape)</i></p>
<p><i>jurukjuruk</i> <i>1. bunyi hujan atau air dsb terus-menerus mengalir dengan cepat dalam jumlah banyak lalu berhenti (the sound of rain or water etc. continues to flow rapidly in large quantities and then stop) 2. atau untuk menyebut bentuk seperti itu (or to refer to such shape)</i></p>	<p><i>jurukjuruk</i> <i>1. gulgeun muljulgina binmul ttawiga ppareuge jakku heureugeona naerineun sori. (the sound of rain or water etc. continues to flow rapidly in large quantities and then stop) 2. ttoneun geu moyang. (or to refer to such shape)</i></p>

Figure 9 above shows the definition of onomatopoeic and mimetic headwords tend to be simple translations of the corresponding definitions provided by the *Standard Korean Language Dictionary*. This translation inaccuracy occurs because it is hard to find their equivalent in Indonesian. Compared to other forms of translation, defining the Korea OM by translating the meaning froms of Standard Korean Language Dictionary is most commonly found in online dictionary.

V. CONCLUSION

The issues regarding the treatment of OM in the Korean–Indonesian dictionary can be classified into four types: 1) POS shifting; 2) inaccurate translation; 3) zero translation; and 4) lack of systematicity. There are 32.3% of translations showing the POS shifting in the printed dictionary and 14.7% in the online dictionary. The tendency of these POS shifting occurs because of some reasons such as difficulty in finding equal words in Indonesian, not only equivalent in meaning but also in POS. On Chun (1996) Indonesian OM root word commonly belong to noun, and can be attached to affixes in the forms of derivational adjective or verb. Therefore the derivational form of OM and the verbs or adjectives followed to OM which are translated, not the OM themselves.

In the printed dictionary, there are 4.4% translation inaccuracies and only 0.4% in the online dictionary. This is likely because in the online dictionary, the translation is done by translating the definitions of OM taken from the Standard Korean Language Dictionary. There are 31 OM, or about 15.3% of all the research data are not found at all in the printed dictionary (zero translation). The absence of OM mentioned above as headword in Korean-Indonesian dictionaries is most likely because in the process of collecting data, those OM words are not selected.

Furthermore, this study found 15% inconsistencies in the printed dictionary and only 1% in the online one. The most commonly found lack of systematicity is related to determining of headwords. Normally, the headword is the root word of the OMs. However, this study found some headwords in the form of derivative words.

Finally, based on the conclusions presented previously, the following suggestions are made: 1) as much as possible the translation must be formulated by finding the Indonesian equal words (OM should be simply translated as OM/interjection) in order to make them easier to understand. If it is hard to find their equivalent in Indonesian, they should be translated according to Standard Korean Language Dictionary. 2) The headword of OM should be put in their root form, followed by their derivative forms (if any), and either its collocating verbs or adjectives must be included as entry words.

This study has some limitations, as it only discusses treatment in translation and determination of the OM headword. There are still many other aspects of dictionary treatment regarding Korean OM that need to be considered for improving the Korean-Indonesian dictionary. In addition, according to Park (2012:4) in Korea there is about 8,283 OM but this study is only limited to 202 data.

REFERENCES

- Chae, Wan. 2003. *Korean Onomatopoeia and Mimesis*. Seoul: Seoul National University Press
- Chun, Tai-Hyun. 1996. "A Study on Iconicity of Indonesian Iterative Method". *Journal of Language and Linguistic* 16: 169-203
- Chun, Tai-Hyun. 2003. "Bilingual Dictionary and Translation: with Special Reference to Indonesian-Korean, Korean-Indonesian Bilingual Dictionary". *Journal of Korealex* 2: 163-190. The Korean Association for Lexicography
- Jung, Yeong-Rim. 1998. *Korean-Indonesian Dictionary: Kamus Bahasa Korea-Indonesia*. Seoul: Hankuk University of Foreign Studies Press
- Naver Korean-Indonesian Dictionary. <https://stdict.korean.go.kr/main/main.do>
- Park, Hye-Won and Bae, Sungbong. 2012. *Korean Onomatopoeic-Mimetic Words*. Ulsan: Ulsan University Press
- Kemmer, Suzanne. "Words in English: Structure". *Words in English*. <http://www.ruf.rice.edu/~kemmer/Words04/structure/index.html>. Retrieved 30 April 2019
- National Agency for Language Development and Cultivation. *Kamus Besar Bahasa Indonesia*. Retrieved from <https://kbbi.kemdikbud.go.id/>
- National Institute of Korean Language. *Korean Standard Dictionary*. Retrieved from <https://stdict.korean.go.kr/main/main.do>

Yoon, Heui-Won. 1993. "Concept and Definition of Onomatopoeic-Mimetic". *Saegugeosenghwal* 3-2, National Insitute of Korean Language

TERMINOLOGY OF ART MUSIC IN OTTOMAN TURKISH AND MODERN TURKISH LEXICOGRAPHY (19TH-21ST CENTURY).

Agata Pawlina

Jagiellonian University

Abstract

In this paper the author presents preliminary conclusions concerning changes in forms (in respect of orthography and morphology) and semantics of musical terms in Ottoman Turkish and modern Turkish lexicography. Selected dictionaries (including mono- and bilingual, general and specialized ones) from the period between the 19th and 21st century had been analyzed. This period of time is particularly interesting for establishing how a) Turkish lexicographic works reflect the Westernization of high-culture music of the late Ottoman Empire and the young Republic of Turkey and b) differences between art music of the Ottoman and European traditions are perceived nowadays.

By presenting a terminological analysis of words considered to be not only basic musicological terms but also a part of natural language ('singer', 'piano', 'kanun') the author unveils some of the issues associated with the translation of Turkish musical terms into European languages and vice-versa. Those problems arise from the duality and hybridity which exist in contemporary Turkish musical culture. Its older part, the so-called classical/traditional art music (tur. *Türk sanat müziği* or *Osmanlı/Türk klasik müziği*) emerged at the turn of the 17th century, as a part of a Middle Eastern art musical tradition. Later on, during modernizing efforts conducted in the declining Ottoman Empire in the 19th century, European art music had been incorporated along with its international, translingual terminology.

As a result of such duality, interesting phenomena are being observed in modern Turkish vocabulary concerning art music. There are "general terms" which can be used in the context of both musical traditions, but there are also "highly-specialized" ones, concerning exclusively Middle Eastern- or Western-style music. That, along with frequent polysemy and a significant number of synonyms, homonyms and homophones, prompts the interpreter of Turkish musical terms to conduct an in-depth investigation of the context in which each term is being used.

Key Words: terminology, lexicography, musical terms, Turkish, Ottoman Turkish

1. Introduction

The main objective of the author's research is to collect, systematize and analyze Ottoman Turkish and modern Turkish specialized vocabulary concerning art music, recorded in lexicographic works. This paper focuses only on dictionaries from the 19th, 20th and 21st century. Along with selected examples of linguistic evidence, preliminary conclusions regarding changes in forms and semantics of Turkish musical terminology will be presented. Therefore, the most important issues associated with the translation of Turkish musical terms into European languages and vice-versa will be explored. By employing the methods of historical lexicography, the author also intends to establish whether the Turkish language, attested in analyzed dictionaries, reflects the Westernization process of art music which took place in Turkey in two stages, different in character and scope: first, in the period of the late Ottoman Empire (1826-1923) and then, after the proclamation of the Republic of Turkey (1923 onwards).

One also has to ascertain that in Turkey, since the *Tanzimat* period (1839 onwards), the language reform – whose primary purpose was to simplify Ottoman Turkish grammar and emancipating its vocabulary from Arabic and Persian borrowings – had been one of the most important topics on the agenda of the political authorities and the intelligentsia (Brendemoen, 1990; Heyd, 1954; Shaw & Shaw, 2012, pp. 214–221; 387–396). Thus, in the Turkish context, a social-cultural approach towards language and its lexicography, utilized by the author in current research, seems especially fitting. As Ewa Siemienieć-Gołaś stated:

“At each stage of the language reform the published dictionaries documented the developments constituting on the one hand the evidence of changes, on the other hand presenting a new image of the language. The dictionaries, their variety and kinds, were not only a reflection of the changes – they became the result of the changes” (Siemienieć-Gołaś, 2015, pp. 141–142).

1.1. Basic terms used

In this paper the author does not wish to elaborate on in-depth musicological issues, yet some basic terms, conceptions and historical processes have to be explained before one presents the results of lexicographical research.

In present article, “art music” or “classical music” (terms which are used interchangeably) is understood as music that a) is professional – performed by musicians educated in specialized institutions; b) is elitist – at some point of its history it was created and performed for and by members of the highest social strata; c) has a well-documented theory of music (Sadie, 2001, pp. 425–437). To simplify the issue, in the popular imagination of Europeans, that description would fit music nowadays performed in concert halls by symphonic orchestras or in opera theaters. However, if we think of contemporary Turkey, the case presents itself as a more complicated one. We can observe a duality in Turkish classical music, in which two completely different genres are being developed independently: Western-style art music and Middle Eastern-style art music, with a hybrid “in-between” niche, consisting of the outcomes of cross-cultural music making.

Middle Eastern art music is a name applied to the great musical tradition of the Arabic-, Persian- and Turkish-speaking world, it can also be called the art music of the Islamic civilization (Danielson, Reynolds, & Marcus, 2002; Faruqi, 1985; Shiloah, 1980, 2001). The “Ottoman idiom” emerged as its youngest stratum, in the second half of the 16th century. At first it had been performed mostly in the palace of the sultan in Constantinople, then reached beyond the palaces of the elite into the urban culture of the Ottoman Empire and to dervish (especially Mevlevi) lodges (Behar, 2006; Feldman, 1996a). In terms of musical theory and performance style, Ottoman-Turkish art music can be characterized as a) monophonic – with a sophisticated system of melodic patterns called *makams* and rhythmic patterns called *usuls*, b) based on unequal-tempered microtonal scale, c) performed, until the 20th century, exclusively for small audiences by soloists or by chamber ensembles, consisted of one or two singers and a few instrumentalists, playing percussion and stringed instruments, d) exhibiting the primacy of vocal music over instrumental music, e) transmitted, until the 19th century, exclusively by the oral tradition (Behar, 1998, 2006; Danielson et al., 2002; Feldman, 1990, 1996b; İhsanoğlu, 2003, pp. XXXI–LI; Karabaşoğlu, 2013; Signell, 2002).

1.2. The Westernization of Turkish musical culture – overview

In 19th century, authorities of declining Ottoman Empire encouraged musicians to adapt Western-style music to make it a symbol of the modernization the army, administration and culture of the country, which started under sultan Mahmud II’s reign (1808-1839) and reached its peak during the *Tanzimat* period (Shaw & Shaw, 2012, pp. 25–415; see also: Aracı, 2006; Komsuoğlu & Turan, 2007; Kutlay

Baydar, 2010; Pawlina, 2014, 2017). The above-mentioned features of the Ottoman Turkish art music could then be conveniently juxtaposed with the characteristics of the 19th-century European classical music: polyphonic, based on equal-tempered scale, performed mostly by great symphony orchestras, for big audiences, in established concert venues or opera theaters and transmitted by well-developed musical notation. Such oversimplified, superficial comparison of Western-style and Middle Eastern-style musical traditions, even though criticized by musicians and scholars of that time, became a part of a wider *Alaturka – Alafranga* dispute – concerning the technological and cultural superiority of the European countries over the declining Ottoman Empire – and led to significant changes in Turkish musical culture (Kaya, 2012; O’Connell, 2000, 2005).

After the proclamation of the Republic of Turkey in 1923, the “musical revolution” changed its character and reached beyond Istanbul. Music is an often neglected field in the context of Kemalist reforms (Shaw & Shaw, 2012, pp. 561–585). However, an examination of sources leads to a conclusion that for Kemal Atatürk and his political advisors, provoking a change in music which Turkish people had been listening to, was no less important than reforms of other aspects of culture and customs (Alpagut, 2011; And, Yener, Altar, & Laszlo, 1982; Ataman, 1991).

Until the death of the first Turkish president in 1938 the “institutional part” of reforms in music had been finished, with new Western-style orchestras and conservatories established (Pawlina, 2018, pp. 24–27). A new generation of composers started to create music in a style recommended by the authorities – a fusion of Western-style art music composition techniques and Turkish folk music (Gökalp, 1968, pp. 129–131; see also: Aracı, 1997; Degirmenci, 2006; Kılıç, 2009; Krone, 1952; Tekelioğlu, 2001).

This new Western-style Turkish art music, back then called *Millî Müsiki* (eng. National Music), along with indigenous Anatolian folk music was promoted by Kemalist authorities for years, while the Ottoman Turkish classical music faced the threat of oblivion. However, as an important part of urban culture it survived and – not unchanged – since 1990s onwards – experiences a period of renaissance (Çolak, 2006; Feldman, 1996a, pp. 16–18; O’Connell, 2005, 2013; Pohlit, 2010; Signell, 1980). Thus, as a result of a Westernization process which occurred in Turkish art music in the 19th and 20th centuries, nowadays, in Turkey two separate genres, Western-style and Middle Eastern-style, are being developed independently.

2. Examined sources and research methods

The author believes that such an extraordinary change in high-culture music of Turkey has to be reflected in the language of each period – the 19th, 20th and 21st century. To verify this hypothesis and to fulfill other research objectives, stated in the Introduction above, lexical material had been excerpted from selected mono- and bilingual general and specialized dictionaries and from music thesauri of the Ottoman Turkish and Turkish languages. The primary sources which were examined are listed here along with an abbreviation (given in [] brackets) which will be used in the tables further below. It is worthwhile to note that for the purposes of this article, the author selected only a few examined sources from a much greater group, which is being used for her current research.

2.1.1. Bilingual dictionaries:

- 1) James W. Redhouse, *A lexicon, English and Turkish: shewing in Turkish, the literal, incidental, figurative, colloquial, and technical significations of the English terms*, London 1861. [R¹⁹]
- 2) Anton B. Tinghir and Kirkor Sinapian, *Dictionnaire français-turc des termes techniques des sciences, des lettres et des arts*, Constantinople 1891. [FT-TS]

3) Fritz Heuser and İlhami Şevket, *Türkisch-deutsches Wörterbuch*, 6th edition, 1967 Wiesbaden; first edition: 1931, Istanbul. [HŞ]

4) *Redhouse Yeni Türkçe-İngilizce Sözlük*, Redhouse Yayınevi, İstanbul 1974. [RY]

5) *SlovoEd Deluxe Turkish-English*, Kindle DX version no 1.5., 2011, database provided by Redhouse [Slo]

2.1.2. Monolingual Turkish dictionaries:

1) Musa Canpolat, (Ed.), *Türkçe Sözlük*, Türk Dil Kurumu, Ankara 1983. [TS]

2) TDK *Büyük Türkçe Sözlük* [BTS], updated version of TS, online: <http://www.tdk.gov.tr>

2.1.3. Thesauri of music:

1) Kâzım Uz, *Musiki istilâhatı*, at first published in Constantinople in 1894; revised, extended and rewritten in Latin script by Gültekin Oransay, partly in modern Turkish, partly in Ottoman Turkish, published in Ankara in 1964. [Uz]

2) Mahmut Ragıp Gazimihâl, *Musiki sözlüğü*, İstanbul 1961. [G]

3) Vural Sözer, *Müzik Ansiklopedik Sözlük*, 5th edition, Remzi Kitabevi, İstanbul 2005. [MAS]

2.2. Methods – Interdisciplinary approach

In order to conduct a terminological analysis of specialized musical vocabulary excerpted from the sources listed above, the author embraces an interdisciplinary approach. The idea for such research was inspired by sociolinguistics, more specifically, the theory of the social-cultural basis of knowledge and its application into lexicographic research (Berger & Luckmann, 1991; Doroszewski, 1970). The methods of historical lexicology were applied to establish the origin of terms and changes in their forms, in respect of orthography and morphology. However, the analysis of the meaning of each term combines the methodology of lexical semantics, including the evaluation of cross-linguistic differences and similarities in lexical-semantic structure, with the results of musicological and historical research regarding the period in which dictionaries had been written.

3. Results

As stated above, in this paper the author does not wish to elaborate on in-depth musicological issues. Therefore, as a tiny illustration of a much greater research result, three basic terms had been chosen: ‘singer’, ‘piano’ and ‘kanun’. In Turkish, all of these terms may be considered not only a part of the specialist musicological lexicon but also as a part of natural language – designations current in everyday speech and literature. Such selection enables the author to reveal some of the most important issues associated with the translation of Turkish music vocabulary into European languages and vice-versa, not only to specialists but also to readers without a musicological background, interested exclusively in the linguistic content of current research.

In the tables below, each term is provided in all forms and meanings found in dictionaries. Sources are indicated by the appropriate abbreviation and are listed in chronological order, from the oldest one. Definitions from general dictionaries are fully quoted, in unchanged orthographical form (neither in reference to Arabic script, nor the modern Turkish alphabet).

Specialized dictionaries [MAS, G, Uz] often present a few pages-long descriptions of concepts. In such a case only the title of the article is presented in the table, with a short summary or commentary on the content of the definition, printed in italics.

The sign ‘>’ means that the source does not describe a term, only points to its synonym.

Table 1. SINGER

source	page(s)	article title	definition(s)
R ¹⁹	687	singer	خواننده
FT-TS	–	–	<i>not included</i>
Uz	29	hanende (hvanende)	Beste veya şarkı okuyan ademe denür. <i>In 1964 edition Oransay added: Irlayıcı.</i>
	34	ırlayıcı	Irlıyan kişi. Eski terimi: muganni/muganniye, hanende, okuyucu, şarkıcı, ses sanatkârı. (Uz: yok) = <i>added by Oransay in 1964</i>
	66	şarkıcı	> ırlayıcı
HŞ	186	hānende	Sānger, Sāngerin.
	420	muganni muganniye	Sānger Sāngerin
	474	okuyucu	1. Leser 2. Sānger
	579	şantöz	(franz. chanteuse) Sāngerin
	491	şarkıcı	Straßensānger
G	–	–	<i>not included</i>
RY	447	hanende; خواننده	Or. mus. [=Oriental music] singer
	788	muganni; مغنى muganniye; مغنيه	singer, male singer professional woman singer in the Arabic style
	898	okuyucu; اوقويغي	1. reader 2. singer 3. one who recites incantations; exorcist 4. person who goes around and invites people to a wedding
	1049	şantöz; شانتوز	female singer
	1050	şarkıcı	1. singer 2. song writer
TS	v.1, 501	hanende	Şarkı söylemeyi meslek edinmiş kimse; şarkıcı, okuyucu.
	v.1, 846	muganni muganniye	Şarkı söylenen kimse; şarkıcı. Şarkıcı kadın.

	v.2, 1109	şantör şantöz	Erkek şarkıcı. Kadın şarkıcı.
	v.1, 1110	şarkıcı	Şarkı söylenen, şarkı söyleme yeteneği olan yada mesleği şarkı söylemek olan kimse; muganni, muganniye.
MAS	326	hanende	Okuyucu, şarkıcı, ses sanatçısı.
	482	muganni	Şarkı söyleyen (erkek), şarkıcı, okuyucu, hanende. Kadın olursa, muganniye.
	513	okuyucu	Ses sanatçısı. Şarkı yada türkü söyleyen kimse.
	671	şantör	Erkek şarkıcı. Ses sanatçısı.
		şantöz	Kadın şarkıcı. Ses sanatçısı
	672	şarkıcı	Şarkı söylenen, mesleği şarkı söylemek olan kimse. Okuyucu.
Slo		hanende	formerly professional singer of Turkish classical music
		şantör	male singer, chanteur
		şantöz	female singer, chanteuse
		şarkıcı	1. professional singer 2. <i>colloq.</i> songwriter
BTS		hanende	<i>esk.</i> [=obsolete] Şarkıcı.
		muganni	<i>same as in TS, but esk. abbreviation had been added</i>
		muganniye	
		şantör	Erkek şarkıcı. Ses sanatçısı.
		şantöz	Kadın şarkıcı. Ses sanatçısı
		şarkıcı	<i>same as in TS, but adds more synonyms: okuyucu, hanende</i>

Table 2. PIANO

source	page(s)	article title	definition(s)
R ¹⁹	576	piano, pianoforte	چمبالو ; پیانو
FT-TS	266	piano	پیانو دینملان هشمور آلت موسیقی
Uz	56	piyano	Alafranga alat-ı musikiyyesinden maruf olan bir alet ismidir.
HŞ	505	piyano	Piano, Klavier

G	205	piano	maruf musiki aletinin adıdır: bunu <i>piyano</i> imlâsile de yazabiliyoruz
	205	piano-forte	(...) İtalya ve İngilterede alete tam yekpare imlâ ile <i>pianoforte</i> dedikleri halde, Fransada ve bizde en az yüz yıldır <i>piyano</i> kısaltması tercih edilegelmiştir.
	206	piyano	<i>organology classification, construction, history of the instrument, the utilization of pianos in music pedagogy, etc.</i>
RY	937	piyano; پیانو	piano
TS	v.2, 966	piyano	Klaviyeli, telli, ağır ve büyük çalgı.
MAS	553-554	piyano	<i>organology classification, detailed construction, history of the instrument, playing techniques, etc.</i>
Slo		piyano	piano, pianoforte
BTS		piyano	Klavyeli, telli, değişik tuşlara basılarak çalınan ağır ve büyük çalgı.

Table 3. KANUN

source	page(s)	article title	definition(s)
R ¹⁹	–	–	<i>not included</i>
FT-TS	–	–	<i>not included</i>
Uz	39	kanun	<i>information about construction and playing techniques</i>
HŞ	257	kānūn	Art Zither
G	–	–	<i>not included</i>
RY	596	kanun; قانون	a zither-like musical instrument with 72 strings
TS	v.1, 639	kanun	Dikdörtgen biçiminde, bir köşesi kesik, yassı bir sandık üzerine gerilmiş tellerden oluşan, tırnak adı verilen çalgıçlarla çalınan incesaz çalgısı.
MAS	381- 382	kanun	<i>information about history, construction and tuning</i>
Slo		kanun (II)	<i>same as RY</i>
BTS		kanun	<i>same as TS</i>

4. Discussion and Conclusions

Observation of lexical evidence, even so limited in number as the one presented above, leads us to the main conclusion: Turkish musical terminology was influenced by both the language reform and the Westernization of art music. Thus, the main hypotheses of the current paper are confirmed. Several processes which occurred in the musical vocabulary in the course of time, inferred from lexicological analysis of content of Table 1., 2. and 3. are briefly characterized below.

4.1. Simplification of terms

By comparing the Ottoman Turkish خواننده [havānende] from R¹⁹ with Oransay's transliteration in the form of 'hanende/hvanende' in Uz and TS 'hanende', we can infer the presence of a tendency, which is natural for the Turkish language, to phonological simplification of terms.

4.2. Polyonymy

An analysis of the 'singer' concept in Table 1. enables us to observe the gradual process of replenishment of the vocabulary with new lexical units. In this and many other cases within musical terminology, it led to the emergence of a significant number of synonyms and near-synonyms. In fact, polyonymy could be regarded as the main issue in contemporary Turkish musical terminology.

In addition to synonyms, we observe in it frequent polysemy, the presence of homonyms and homophones and the phenomena sometimes referred to as the "false friends" of the translator. The constraints of the current paper do not allow the author to present examples of all of those phenomena, yet it is worthwhile to note that Turkish musical vocabulary requires further systematization and standardization to avoid the inevitable ambiguity induced by them.

4.3.1. Influences of the Turkish language reform – orthography

At least two remnants of the Turkish language reform are reflected in Table 1. The obvious one is the transformation of the written forms of terms from the Arabic to the Latin script. A comparison of those forms between R¹⁹, FT-TS and HŞ, RY, TS, BTS enables us to observe not only the change of the alphabet itself but also post-1928 changes in the attitude towards modern Turkish orthography, especially in terms of indicating long vowels or otherwise.

4.3.2. Influences of the Turkish language reform – nativization of vocabulary The second remnant of the language reform is the creation of new designations for the same concept with the purpose of emancipating Turkish vocabulary from Arabic and Persian borrowings. By looking at Table 1 we may conclude that the goal of Turkification of the term 'singer' had been achieved. Currently [MAS, Slo, BTS] the term 'şarkıcı' seems to be most common in everyday speech. The Persian form 'hanende' and the Arabic form 'muganni/muganniye' are known mostly to specialists and performers of the Ottoman Turkish art music. Distribution of the most recent loanwords, the French 'şantör/şantöz', is also limited.

It is necessary to conduct statistical analysis and further research in the field of contextology to draw specific conclusions regarding the distribution of each synonym which occurs in Turkish musical terminology. However, some of the assumptions are presented in section 4.4.1.1. below.

4.4.1. Influences of Westernization – two subclasses in vocabulary

Due to the cultural duality of contemporary Turkish art music, we can differentiate two main groups in musical vocabulary. There are "general terms" which can be used in the context of both Western- and Middle Eastern musical traditions and "highly-specialized terms" concerning exclusively Middle Eastern- or Western-style music.

4.4.1.1. General terms

The first group encompasses e.g.: “objective” acoustic phenomena, such as ‘pitch’, ‘sound’, ‘tempo’; terms which refer musical notation, e.g. ‘note’, ‘staff’, ‘flat’, ‘sharp’ or terms which ascribe a whole range of musicological classifications, for instance in organology (‘wind instruments’, ‘stringed instruments’, etc.) or the history and theory of music (‘form’, ‘composer’, ‘genre’, etc.).

If we think of speaking about Turkish music in the English language, the word ‘singer’ can be considered as an example of this group of terms because it can be used in reference to all musical genres (popular, folk, art – both Western-style and Middle Eastern-style). Yet, as we can observe in Table 1, that is not the case in the Turkish language, which developed in the course of time many distinctive units designating – ostensibly – the same concept of a ‘person who sings’. Based on the content of definitions we can assume that in everyday speech the Turkish word ‘şarkıcı’ is the most commonly used term to refer to such a person, regardless of his or her gender, whether it is his or her job or hobby, or which musical genre he or she performs. But in the case of specialist discourse, during the translation from e.g. English to Turkish, one should investigate further: who is singing and what is being sung – an opera aria, pop song, *ilâhi* (a religious form in the Middle Eastern-style art music) or a *kâr* (a secular form in the Middle Eastern-style art music) and it is only afterwards that one should choose the proper term.

4.4.1.2. Highly-specialized terms

As representatives of the second group in musical terminology, two names of instruments had been chosen (Table 2. and 3.) because this semantic field is the easiest to comprehend without in-depth musicological background. ‘Piano’ serves as an example of vocabulary concerning exclusively Western-style classical music. On the other hand, ‘kanun’ (in English also spelled ‘qanun’) is a name of a stringed instrument used in the Ottoman Turkish art music. Other fields that may be listed as “highly-specialized terms” include e.g. names of Turkish *makams* and *usuls*, names of genres, performing styles, composing techniques, etc.

4.5. Influences of the “musical revolution” (1923-1938)

Gâzimiha’s thesaurus of music [G] is a very important source for current research not only because it was the first modern work of this type. More importantly, it seems to be highly influenced by early Kemalist ideas on Turkish music. According to the principles of the republican “musical revolution”, Ottoman Turkish art music should not be performed and developed anymore and Turkish people should practice only their indigenous folk music and its fusion with Western-style art music. It seems to be the reason why in Table 2. and 3. we may observe three designations for ‘piano’ but ‘kanun’ – one of the most important instruments of the Ottoman Turkish art music – was not included in the dictionary. In fact, G does not provide terms designating Middle Eastern-style art music in the form of article titles; they are only mentioned several times in the content of definitions of some other concepts. On the other hand, one can find in it a great deal of terms regarding Turkish folk music and European classical music.

4.6. Musical vocabulary as an indicator of cultural change

The author believes that the presence or absence of some “highly-specialized” terms in dictionaries written in the 19th and 20th century may be considered as an indicator of the pace of the process of westernization in the Ottoman Turkish and post-Republic Turkish culture. This topic requires further study, but in order to exemplify the issue the author wishes to present an example of such deduction based on the content of Table 2 and 3.

When Kâzım Uz was writing his dictionary (1864) the piano was already a part of the Ottoman Turkish musical culture (at least since 1827 when sultan Mahmut II brought this instrument to his palace), but it seems it was still considered foreign (“Alafranga”). This could also be attested by a long definition of the new instrument, maybe still unknown to some of the readers of the dictionary, presented by Tinghir-

Sinapian. Interestingly, Redhouse gives two translations: پیانو [pīāno]; چمبالو [çembālo], of which the second actually designates another instrument, the harpsichord.

100 years later, Gâzimiha1 presents three types of the spelling of the term ‘piano’. Two of them, ‘piano’ and ‘piano-forte’ are given only to introduce terminological recommendations to use only the third, the “Turkish” form – ‘piyano’. Under the latter, he describes the European genesis of the instrument (he also mentions the harpsichord as the “ancestor” of the piano) with some remarks on how it had been incorporated into Turkish culture in the 19th century. The word ‘alafranga’, which indicates foreign provenance of piano, disappeared from this and from later descriptions of the term.

Taking everything which was stated above into consideration, a comparative evaluation of dictionaries written in Turkey in the 19th, 20th and 21st-century and an analysis of the musical terminology attested in them may constitute a valuable source of information for researchers interested not only in Turkish lexicography, linguistics and musicology, but also for historians who explore the field of cultural change in the declining Ottoman Empire and the young Republic of Turkey. However, historical lexicographic research should be considered as a preliminary stage of much wider future research. In order to explore the field in a comprehensive manner in the context of contemporary Turkish culture, an investigation which employs pragmatics, contextology and statistical methods of analysis is necessary.

References

1. Dictionaries

- Canpolat, M. (Ed.). (1983). *Türkçe Sözlük*. Ankara: TDK.
- Gazimihâl, M. R. (1961). *Musiki sözlüğü*. İstanbul: Millî Eğitim Basımevi.
- Heuser, F., & Şevket, İ. (1967). *Türkisch-deutsches Wörterbuch* (6th ed.). Wiesbaden: O. Harrassowitz.
- Redhouse, J. W. (1861). *A lexicon, English and Turkish shewing in Turkish, the literal, incidental, figurative, colloquial, and technical significations of the English terms*, London: Oriental Literature Society.
- Redhouse Yeni Türkçe-İngilizce Sözlük*. (1974). İstanbul: Redhouse Yayınevi.
- SlovoEd Deluxe Turkish-English Dictionary* (Kindle DX edition). (2011). Retrived from Amazon.com.
- TDK Büyük Türkçe Sözlük online*. (b.d.). Retrived from <http://www.tdk.org.tr>
- Tinghir, A. B., & Sinapian, K. (1891). *Dictionnaire français-turc des termes techniques des sciences, des lettres et des arts*. Constantinople: Impr. de K. Bagdadlian.
- Uz, K. (1964). *Musiki istilâhatı* (G. Oransay, Ed.). Ankara: Küğ.
- Sözer, V. (2005). *Müzik Ansiklopedik Sözlük*. İstanbul: Remzi Kitabevi.

2. Other publications

- Alpagut, U. (2011). *Müzik Sorunlarına Bakışta Atatürk'ün İzleri*. Oxford.
- And, M., Yener, F., Altar, C. M., & Laszlo, F. A. (1982). *Atatürk Türkiyesi'nde Müzik Reformu Yılları*. Münich.
- Aracı, E. (1997). Reforming Zeal. *The Musical Times*, 138(1855), 12–15.
- Aracı, E. (2006). *Donizetti Paşa: Osmanlı sarayının İtalyan maestrosu*. İstanbul: YKY.
- Ataman, S. Y. (1991). *Atatürk ve Türk musikisi*. Ankara: Kültür Bakanlığı Yayınları.
- Behar, C. (1998). *Aşk olmayınca meşk olmaz: geleneksel Osmanlı/Türk müziğinde öğretim ve intikal* (1st ed.). İstanbul: Yapı Kredi Kültür Sanat Yayıncılık.
- Behar, C. (2006). The Ottoman Musical Tradition. In S. Faroqhi (Ed.), *The Cambridge History of Turkey* (Vol. 3, pp. 398–407). New York: Cambridge University Press.
- Berger, P. L., & Luckmann, T. (1991). *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*. London: Penguin Books.
- Brendemoen, B. (1990). The Turkish Language Reform and Language Policy in Turkey. In G. Hazai (Ed.), *Handbuch der türkischen Sprachwissenschaft I* (pp. 454–493). Budapest.
- Çolak, Y. (2006). Ottomanism vs. Kemalism: Collective Memory and Cultural Pluralism in 1990s Turkey. *Middle Eastern Studies*, 42(4), 587–602.
- Danielson, V., Reynolds, D., & Marcus, S. (Eds.). (2002). *The Garland Encyclopedia of World Music. Volume 6, The Middle East* (Vol. 6). New York, London: Routledge.
- Degirmenci, K. (2006). On the Pursuit of a Nation: The Construction of Folk and Folk Music in the Founding Decades of the Turkish Republic. *International Review of the Aesthetics and Sociology of Music*, 37(1), 47–65.

- Doroszewski, W. (1970). *Elementy leksykologii i semiotyki*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Faruqi, L. I. al. (1985). Structural Segments in the Islamic Arts: The Musical ‘Translation’ of a Characteristic of the Literary and Visual Arts. *Asian Music*, 16(1), 59–82.
- Feldman, W. (1990). Cultural Authority and Authenticity in the Turkish Repertoire. *Asian Music*, 22(1), 73–111.
- Feldman, W. (1996a). *Music of the Ottoman Court: makam, composition and the early Ottoman instrumental repertoire*. Berlin: VWB-Verlag für Wissenschaft und Bildung.
- Gökalp, Z. (1968). *Türkçülüğün Esasları* (7th ed.). Istanbul: Varlık Yayınları.
- Heyd, U. (1954). *Language Reform in Modern Turkey*. Jerusalem.
- İhsanoğlu, E. (2003). *Osmanlı müsiki literatürü tarihi*. İstanbul: İslâm Tarih, Sanat ve Kültür Araştırma Merkezi IRCICA.
- Jin-Ah Kim. (2017). »Cross-Cultural Music Making«: Concepts, Conditions and Perspectives. *International Review of the Aesthetics and Sociology of Music*, 48(1), 19–32.
- Kapchan, D. A., & Strong, P. T. (1999). Theorizing the Hybrid. *The Journal of American Folklore*, 112(445), 239.
- Karabaşoğlu, C. (2013). Meshk: As a Tradational Method of Turkish Music Education. *Procedia - Social and Behavioral Sciences*, 106, 1834–1839.
- Kaya, E. E. (2012). Yeni Türk Müzik İnkılâbına Bir ‘Hazırlık Evresi’ Olarak 1826-1920 Dönemi. *Turkish Studies - International Periodical For The Languages, Literature and History of Turkish or Turkic*, 7(1), 1451–1460.
- Kılıç, F. (2009). Çok Sesli Batı Müziğinin Türk Modernleşmesindeki Önemi. In 38. ICANAS (Uluslararası Asya ve Kuzey Afrika Çalışmaları Kongresi) 10-15.09.2007, Ankara, Bildirler: 13. Müzik kültürü ve eğitimi (Vol. 1, pp. 455–464).
- Komsuoğlu, A., & Turan, N. S. (2007). From Empire to the Republic: the Western Music Tradition and the Perception of Opera. *International Journal of Turcologia*, 2(3), 5–29.
- Krone, M. T. (1952). Music in Turkey. *Music Educators Journal*, 39(2), 28–30.
- Kutlay Baydar, E. (2010). Osmanlıda Görevli İki İtalyan Müzisyen: Giuseppe Donizetti ve Callisto Guatelli. *Zeitschrift Für Die Welt Der Türken / Journal of World of Turks*, 2(1), 283–293.
- O’Connell, J. M. (2000). Fine Art, Fine Music: Controlling Turkish Taste at the Fine Arts Academy in 1926. *Yearbook for Traditional Music*, 32, 117–142.
- O’Connell, J. M. (2005). In the Time of Alaturka: Identifying Difference in Musical Discourse. *Ethnomusicology*, 49(2), 177–205.
- O’Connell, J. M. (2013). *Alaturka: style in Turkish music (1923-1938)*. UK: Ashgate.
- Pawlina, A. (2014). Muzyka klasyczna Europy w Imperium Osmańskim. *Przegląd Orientalistyczny*, 1–2, 61–76.
- Pawlina, A. (2017). Dźwięki Zachodu na Wschodzie - westernizacja kultury muzycznej Turcji w XIX i XX wieku. *Między Wschodem a Zachodem, Między Północą a Południem*, 259–267. Warszawa: Campidoglio.

- Pawlina, A. (2018). Turkizm w muzyce. Związki muzyki i polityki w młodej Republice Tureckiej (1923-1938). *Wrocławskie Studia Erazmiańskie*, 12, "Orient Daleki i Bliski", 15–33.
- Pohlit, S. (2010, 26.09). *Musical Life and Westernization in the Republic of Turkey*. Presented at the Europa in Opera – Musical Composition of an Identity, Porto. Retrieved from <http://www.celsius-europe.eu/?cat=8>
- Sadie, S. (Ed.). (2001). *The New Grove Dictionary of Music and Musicians*. New York, London: New York Grove, London MacMillan.
- Shaw, S. J., & Shaw, E. K. (2012). *Historia Imperium Osmańskiego i Republiki Tureckiej* (Vol. 2; B. Świetlik, Trans.). Warszawa. [Polish translation of *History of the Ottoman Empire and Modern Turkey*, Cambridge 1977].
- Shiloah, A. (1980). The Status of Traditional Art Music in Muslim Nations. *Asian Music*, 12(1),
- Shiloah, A. (2001). *Music in the World of Islam: A Socio-Cultural Study* (Reprint edition). Detroit: Wayne State University Press.
- Siemieniec-Golaś, E. (2015). Some Remarks on Turkish Dictionaries Published in Constantinople/Istanbul Before and Soon After Language Reform in Turkey (1928). *Rocznik Orientalistyczny*, LXVIII(2), 134–142.
- Signell, K. (1980). Turkey's Classical Music, a Class Symbol. *Asian Music*, 12(1), 164–169.
- Signell, K. (2002). Contemporary Turkish Makam Practice. In V. Danielson, D. Reynolds, & S. Marcus (Eds.), *The Garland Encyclopedia of World Music. Volume 6, The Middle East* (Vol. 6, pp. 47–58). New York, London: Routledge.
- Sutton, R. A. (2013). Musical Genre and Hybridity in Indonesia: Simponi Kecapi and Campur Sari. *Asian Music*, 44(2), 81–94.
- Tekelioğlu, O. (2001). Modernizing Reforms and Turkish Music in the 1930s. *Turkish Studies*, 2(1), 93–108.

SWITCHING FROM ARABIC LEXICOGRAPHICAL TRADITION TO RUSSIAN: CASE STUDY – TATAR DICTIONARIES

Alina Minsafina

Istanbul University

Abstract

The aim of this paper is to point out the differences in Tatar dictionary structure which occur after Tatars gave up Arabic linguistic tradition and started following a Russian one. For this purpose, the following monolingual dictionaries were chosen: *Lehzhe-i Tatari* by Kayyum Nasyri as an example of Arabic tradition and two issues of *Explanatory Dictionary of Tatar Language* (1977-1981 and 2015-2017) as an example of Russian tradition.

After adopting Islam in 922, Tatars were closely connected to Arab culture and science, while the theory of inimitability of Koran raised the need to learn Arabic. It is well known that dictionaries were the greatest achievement of Arab linguistics. The history of Tatar monolingual lexicography starts with the work of Kayyum Nasyri titled *Lehzhe-i Tatari*. The structure of the dictionary, explanation strategy of entry words and their organization according to the Arab alphabet shows the adherence to the principles of Arabic lexicography. While being a part of Soviet Union, Tatars were under the influence of Russian language planning, so Russian became more preferable than Arabic. The revolutionary moment in Russian dictionaries came with the publication of Dmitry Ushakov's *Explanatory Dictionary of Russian Language*, where a new approach to dictionary compiling was given, especially evident in microstructure of the dictionary. This influenced monolingual lexicography of Tatar language. The principles developed in Russian lexicography were applied to *Explanatory Dictionary of Tatar Language* first published in 1977-1981 and modern *Explanatory Dictionary of Tatar Language* with first issue in 2015.

Key Words: Arabic lexicographical tradition, Russian lexicographical tradition, Tatar monolingual dictionaries

Introduction

Throughout history, Tatar people as well as their culture and language, have experienced the influence of both the West and the East. The influence of the East chronologically was earlier.

At the beginning of the 10th century on the territory of modern Tatarstan⁸ monotheistic Islam gradually became the dominant religion. After the visit of Arab missionary Ibn Fadlan to Volga Bulgaria in 922 the Bulgars converted to Islam. The adoption of Islam led to a strong and long-lasting connection of the Tatar nation to the Arab world. In the Volga region Arab-Muslim culture started to form and later influenced social, economic, and political atmosphere in the region. The most powerful was the introduction of religion and Arab language. Islam as a basis of spiritual life of the state enriched the region with new mutual influences, contacts, and synthesis of Turkic and Arab cultures. Tatars started being a part of Arab-Muslim continuum and actively used the achievements of flourishing Arab science. As regards the Tatar language it also experienced the influence of Arabic. A large number of borrowings

⁸ In the 10th century on the territory of modern Tatarstan there was a historical country called Volga Bulgaria.

from Arabic connected with religion, mathematics, astrology, linguistics terminology entered the Tatar language. There are two main factors which played an important role in introduction of Tatars to Arab-Muslim culture. The first one is learning Arabic. The main reason for learning Arab language was the need to read and understand Quran (a holy book of Muslims written in Arabic). Since the revelations of Allah were to be preserved in their original form, the text of Quran should not be translated into any language. In addition, all religious sources because of appearing on the territories where the Arabs lived, were without any doubt in Arabic. Besides Arabic also gave access to scientific heritage of Arab-Muslim world. The second factor is the Arabic graphics. Although the sign system of Arabic alphabet is not suitable to express the phonetics of Tatar language, this type of writing has been in use for almost 1000 years. The acquisition by Bulgars of Arabic writing system contributed to the transition from oral form of creativity to the written one, which allowed to leave to the descendants creative legacy.

From the second part of the 19th century modern Tatar language started its formation on the basis of Kazan dialect. The process of formation ended in the early 20th century. In reforming the Tatar language, at the first step the main role belonged to Kayyum Nasyri. It was he who created the first explanatory dictionary of Tatar language titled *Lehzhe-i Tatari*. Though the main reason for compiling the dictionary was to get rid of Arabic influences in Tatar language, the dictionary itself was based on the principles of Arabic lexicography.

After the revolution of 1905-1907 the situation in the field of reforming the Tatar language has changed dramatically: The convergence of literary language with the spoken language is observed. After the territory of modern Tatarstan entered USSR, Tatar language became a part of language construction. The first factor that is needed to be highlighted in this context is the transition of Tatar writing to the Cyrillic alphabet in 1939 after a short period of using the Latin alphabet. Besides terminology in Tatar language commenced being developed, first based on actual Tatar and Arabic-Persian vocabulary, later starting with 1930s based on Russian and international words. The Tatar language was influenced by language politics of USSR, therefore all linguistic developments from this period are connected with Russian linguistics achievements which last till nowadays.

Methods

A comparative and descriptive method has been chosen for this research. First *Lehzhe-i Tatari* was described as a representative of Arabic lexicographical tradition. Therefore, some common information on Arabic lexicography is given. After that *Explanatory Dictionary of Tatar Language* (1977-1981 and 2015-2017) as an example of Russian lexicographical tradition is described. For this purpose, common information on Russian monolingual dictionaries is given.

Arabic lexicographical tradition

Rooted in Arabic-Islamic culture and firmly linked to other indigenous scholarly interests, particularly Quranic sciences and grammar, the lexicographical tradition represents a major aspect of the Arabic linguistic heritage. In Middle Ages in Arabic sciences word was the topic of research of two main branches: *ilm al-luga* (a domain that explores meanings of words, dialectical variations, strange usage (garib)) and *ilm an-nahw* (which deals with syntax, morphology/morphophonology and, to a lesser extent, phonetics). The early distinction between *nahw* and *luga* eventually gave rise to two diverse, but related, traditions, namely, grammar and lexicography. Lexicography, as well as grammar, had an immense influence on Arabic culture, which is best demonstrated by the fact that the “classical” period of each extends over the thousand years (Baalbaki, 2014, VII, Рыбалкин, 1990, 8).

Belkin (1975, 164) points out that Arabic lexicography originated at the time when this kind of linguistic activity was not familiar with Europe. Muslim world has an ideal environment for the birth and later development of lexicography. Belkin shows several reasons for this, such as the establishment of Arabic

literary language as standard language and as a language of Middle Age Muslim East, the flourishing of Arabic literature and the usage of Arabic as an intermediary language in the scientific activities of other non-Arabic Muslim people. Besides the need to maintain a high level of the standard written language and the spread of dialects of Arabic as spoken languages resulted in accumulation of distinctive features of spoken and written variants of this language which also contributed in Arabic lexicography origin and its later prosperity for many years.

In Rybalkin's opinion (1990, 10) it is most likely that Arab owed the idea of the dictionary to Alexandrian Hellenistic school. The techniques of working with lexical material probably could have come from Sanskrit lexicography. However, it is not possible to find out the ways and the time when lexicographical ideas from Ancient Greece or Ancient Indian culture had entered the early medieval Arab science. Neither Greek, nor Sanskrit or Arabic sources have any positive evidence to prove this idea.

Based on the introduction of Ibn Sida to *al-Muhassas*, Baalbaki (2014, 402) divides Arabic dictionaries to two types: *mubawwab* and *mugannas*. These terms have been adopted to refer to the onomasiological and semasiological types respectively. The first type, in which meaning leads to a sign, mostly comprises a vast number of specialized dictionaries that deal with a specific theme, in addition to a few multithematic works or thesauri. On the other hand, general unspecialized lexica in which sign leads to meaning and which represents the *mugannas* type are considered by fewer in number but posted for their authors more serious challenges, not only in the arrangement of lexical items, but also in the internal arrangement of the lemmata and the extend of the corpus to be included.

Belkin (1975, 165) shows a similar classification of Arabic dictionaries. He points out that among Arabic monolingual dictionaries two types of dictionaries are presented. They are 1) classifying dictionaries in which vocabulary is systemized on conceptual basis (ideological, or thematic dictionaries, or dictionaries arranged according to specific lexical or semantic basis) and 2) explanatory dictionaries where all the words of language are tried to be presented.

If we take into account that the above classification of Arabic dictionaries is based on the feature of "content of the dictionary", it should be also pointed out that there is a classification of dictionaries based on the arrangement of roots. The tendency to this classification is observed in early researches of Arabic lexicography. Rybalkin (1990, 11) notes that different authors mainly distinguish the following types of Arabic dictionaries: 1) dictionaries that use anagram principle Haywood 1975, 524-525; Belkin 1963 222-226; Naji 1971, 377; Darvish 1956, Akhvlediani 1981, 91); 2) dictionaries that use rhymed principle (Haywood 1975, 524-525, Naji 1971, 377; Belkin 1963, 222-226, Akhvlediani 1981, 91); and 3) dictionaries that use alphabetic order (Bielawski 1970, 249-251, 1970a, V; Belkin 1963, 222-226, Akhvlediani, 1981, 91). This classification is based on the way the roots in the dictionary are arranged, as Arabic lexicographers paid attention to root arrangement more that to later arrangement of word derived from these roots.

In our study explanatory dictionaries with alphabetic order of lemmata arrangement are the case as *Lehzhe-i Tatari* is an explanatory dictionary of Tatar language where entries are arranged according to alphabet order.

Russian lexicographical tradition

The beginning of Russian lexicography dates back to the 11th century. First dictionaries called *lexicon*, *alfavit* (= alphabet) or *tolkovanie* (= explanation) were mostly collections of foreign and outdated words. The first dictionary which reached our days is *Kormchaya kniga* (1282), in which the explanations of 174 old Russian, Greece and Old Slavic word are given.

An important dictionary of Russian was created by Vladimir Dal. It is Dal who actually was the first to use the term “explanatory” for dictionaries describing words of the mother tongue.

The new stage in Russian lexicography begins with the publication of *Explanatory Dictionary of Russian Language* edited by D. Ushakov. The main purpose of compilers was to create a dictionary which would reflect both the richness of the language and the changes occurred in the Russian language after post-October period.

The dictionary was aiming to show the whole commonly used the vocabulary of the language from the point of view of its norms and give systematic connections of lexical units. The normativity of the dictionary is manifested in consistent characterization of grammatical properties of the word, its pronunciation and normative stress and carefully developed system of stylistics notes. The system of stylistic notes proposed by this dictionary formed the basis of stylistic characteristics of a word in later explanatory dictionaries of Russian language. The sources of the dictionary were works of literature, social and political journalism, scientific works. For the first time in lexicographical practice the illustration material was extracted from the work of literature.

In this research, the *Explanatory Dictionary of Tatar Language* printed in 1977-1981 and proceeding edition of *Explanatory Dictionary of Tatar Language*⁹ are analyzed from the point of view of their macro- and microstructures.

Results

The results of analyzing three explanatory dictionaries of Tatar language are shown in the table below. The table contains the three explanatory dictionaries of Tatar language issued in different periods of time. Year of publishing, writing system, macrostructure and microstructure of the dictionaries are selected as main parameters to describe the dictionaries.

Table 1 Comparative analysis of three explanatory dictionaries of Tatar language

Parameters	Lehze-i Tatari	Explanatory dictionary of Tatar language 1973-1981	Explanatory dictionary of Tatar language 2015-...
Published in	1 st volume – 1895 (ص - آ) 2 nd volume – 1896 (ض - ي)	1 st volume – 1977 (А-Ӣ) 2 nd volume – 1979 (К-С) 3 rd volume – 1981 (Т – һ)	1 st volume – 2015 (А-В) 2 nd volume – 2016 (Г-Ӣ) 3 rd volume – 2017 (К) 4 th volume - ...
Alphabet used in dictionary	Arabic	Cyrillic	Cyrillic

Macrostructure of a dictionary:

⁹ The first volume of this dictionary was published in 2015 and contains first three letter of Tatar alphabet (А-В), the second volume was published in 2016, the third volume published in 2017 stops on letter К.

1.Introduction	+	+	+
2.Sources used for dictionary compiling	-	1.the list of dictionaries used during the compilation	1.linguistic sources used during the compilation
		2.information on other sources used during the compilation	2. the list of dictionaries used during the compilation
			3. information on other sources used during the compilation
3.How to use the dictionary	-	1.strategy of collecting entries	1.strategy of collecting entries
		2.the structure of the dictionary	2.the structure of the dictionary
		3.strategies to explain the meaning of entries	3.strategies to explain the meaning of entries
		4.representation of phraseological units	4.representation of illustrative material
		5.grammatical and stylistic characteristics of the entries	5. techniques to show borrowings
		6. representation of illustrative material	6.representation of parts of speech
		7. representation of orthography and stress	7. representation of stress

4. Abbreviations used in the dictionary	-	+	+
5. Alphabet order	-	Tatar alphabet based on Cyrillic	Tatar alphabet based on Cyrillic

Microstructure of a dictionary:

1. entries	Every new entry is distinguished only by a new paragraph	Every new entry is written with capital letters, bold type and starts with a new paragraph	Every new entry is written with capital letters, bold type and starts with a new paragraph
2. characteristics of the entry	In some cases, the origin language of the borrowing is shown	Part of speech and stylistic characteristics of the entry is shown	Part of speech, language of origin and stylistic characteristics of the entry is shown
3. explanation of a meaning	In most cases synonyms of the entry are used to explain the meaning; more rarely – antonyms. All the meanings are given in the same article, the polysemy or homonymy of the entry is not distinguished.	All the meanings used in Tatar shown. The most common meaning comes first. Polysemy is shown by Roman numerals, starts with new paragraph. Homonymy is shown by Arab numerals	All the meanings used in Tatar shown. The most common meanings come first. Every new meaning starts with new paragraph. Polysemy is shown by Roman numerals, starts with new paragraph. Homonymy is shown by Arab numerals.
4. illustrative material	In some articles word-combinations for the entry are shown	The usage of the meaning is illustrated by the example sentence. The sentences are selected from works of literature, social and political journalism,	The usage of the meaning is illustrated by the example sentence. The sentences are selected from works of literature, social and political journalism,

scientific works, folklore, and textbooks. Next to the example the origin source of the sentence is given. In some cases illustrative material consists of a word-combination, not a full sentence.

5.phraseological units	In some articles word-combinations for the entry are shown	Phrasal verbs and other phraseological units are given in the article, start with new paragraph after ◇ sign	Phrasal verbs and other phraseological units are given in the article, start with new paragraph after ◇ sign
------------------------	--	--	--

Discussion and Conclusion

Our results led us the following conclusions:

1. Tatar language from the time Tatars adopted Islam was under the influence of Arab-Muslim culture. After becoming a part of the Russian state and later USSR Tatar language was influenced by Russian language. These influences can be observed in language as adopting different writing systems: Arabic alphabet during Arab-Muslim period and Cyrillic alphabet during Russian influence period.
2. The influence of two cultures can be also observed in lexicography. *Lehzhe-i Tatari* compiled according to Arabic lexicographical tradition. The dictionary is an explanatory dictionary arranged to Arab alphabetic order. It gives an explanation of Arabic and Persian words in Tatar, Arabic or Persian variants of Tatar words. In most cases words of different languages are given as explanation of the entry, so the most common way of giving the meaning is the list of its synonyms, more rarely antonyms. As illustrative material some word-combinations are given. Every entry is given with new paragraphs. The differentiation of polysemy and homonymy is not represented. The macrostructure of the dictionary is not complicated and represented only by introduction.
3. *Explanatory Dictionary of Tatar Language* issued in 1977-1981 and the new edition of *Explanatory Dictionary of Tatar Language* started in 2015 shows adherence to Russian lexicographical tradition of monolingual dictionaries which was developed in *Explanatory Dictionary of Russian Language* edited by D. Ushakov. In explanatory dictionaries of this period we can see advanced macrostructure of the dictionary which includes introduction to the dictionary, sources used during the compiling of the dictionaries, the guidance to dictionary usage, abbreviations used in dictionary articles and Tatar alphabet order based on Cyrillic.
4. The microstructure of the dictionary also differs from *Lehzhe-i Tatari*. The article in *Lehzhe-i Tatari* is limited by entry, its meaning, sometimes notes on language origin of the entry, and word-combination as illustrative material. The entries in *Lehzhe-i Tatari* are distinguished only by new paragraphs which makes it harder to read or find the necessary entry.
5. In new dictionaries of Tatar language the advanced system of labels is used. They show grammatical, stylistic and language origin characteristics (*in Dictionary of 2015*) of the entry. In contrast to *Lehzhe-*

i Tatari new explanatory dictionaries separate polysemy (by Rome numerals and new vocabulary entry) and homonymy (by Arab numerals).

6. Illustrative material in new explanatory dictionaries is much richer and selected from different current sources. Phraseological units are also distinguished by new paragraph and \diamond sign.
7. In common, it can be said that both the macrostructure and microstructure of the dictionaries had been developed under the influence of Russian linguistics achievements. New explanatory dictionaries are easier to use and find necessary data. *Lehzhe-i Tatari* though remains as an important seminal monument of the late 19th century which can be used to research the peculiarities of Tatar language of that period.

References

Dubichinckiy, V. (1988). *Theoretical and Practical Lexicography*. Vienna-Kharkov.

Haywood, J. A. (1965). *Arabic Lexicography*. Leiden: Brill

Ramzi, B. (2014). *The Arabic Lexicographical Tradition: from 2nd /8th to the 12th /18th century*. Leiden-Boston: Brill.

Белкин, В. М. (1975). *Арабская лексикология*. М.: МГУ.

Козырев В. А., Черняк В. Д. (2015). *Лексикография русского языка: век нынешний и век минувший*. СПб.: Изд-во РГПУ им. А. И. Герцена.

Насыри, А. (1895). *Ләһжеи Татари Жилди әүвәл*. Казан: Типо-литография Императорского Университета.

Насыри, А. (1896). *Ләһжеи Татари Жилди сани*. Казан: Типо-литография Императорского Университета.

Рыбалкин, В. С. (1990). *Арабская лексикографическая традиция*. Киев: Наук. Думка.

Татар теленең аңлатмалы сүзлеге: I том: А-В. (2015). Казан: ТӘҺСИ.

Татар теленең аңлатмалы сүзлеге: I том: Г-Й. (2016). Казан: ТӘҺСИ.

Татар теленең аңлатмалы сүзлеге: I том: К. (2017). Казан: ТӘҺСИ.

Татар теленең аңлатмалы сүзлеге. Өч томда, I том (1977). Казан: Татарстан китап нәшрияты.

Татар теленең аңлатмалы сүзлеге. Өч томда, II том (1979). Казан: Татарстан китап нәшрияты.

Татар теленең аңлатмалы сүзлеге. Өч томда, III том (1981). Казан: Татарстан китап нәшрияты.

Ушаков, Д. Н. (2013). *Толковый словарь современного русского языка*. М.: Аделант.

TURKISH LEARNER’S DICTIONARY: A NEED OR A LUXURY?

Anna Golynskaia

Istanbul University Cerrahpaşa

Abstract

Teaching Turkish as a foreign language is gaining momentum both inside and outside Turkey. The number of textbooks published in this field is increasing with every passing day. However, one cannot fail to mention the lack of a learner’s dictionary, which is one of the most important tools that help learners to develop learner autonomy. The goal of this study is to develop a model of a monolingual dictionary aimed at those who learn Turkish as a foreign language and find out if using of the given dictionary has any impact on the receptive and productive skills of learners. For this purpose, several words and idiomatic expressions that could be taught at the Upper-Intermediate level were selected. The entries of the selected lexemes were prepared based on the data of the Turkish National Corpus and the Sketch Engine. Then an experiment was conducted. The participant group comprised 31 Upper-Intermediate (B2) Turkish learners from twenty different countries. The choice of the B2 level can be explained by the fact that it is at this level that the teaching of grammatical structures used in the definitions takes place. The subjects had to fill out a questionnaire including six demographic questions and three factual questions concerning dictionary use and fulfil a number of decoding and encoding tasks. The experiment showed that not all of its participants met the stated level of language proficiency, which made it impossible to draw a conclusion regarding either efficiency or inefficiency of using a monolingual Turkish learner’s dictionary.

Key Words: learner’s lexicography, teaching Turkish as a foreign language, monolingual dictionaries, needs analyses

INTRODUCTION

As is known, modern Turkish is quite a young language. Before Atatürk’s Reforms Turkish had incorporated a great deal of Arabic and Persian words and grammar constructions, and it had been written using a Turkish form of the Perso-Arabic script. The adoption of the Latin script in 1928 followed by the purification of vocabulary brought about a completely new language which needed a new teaching strategy.

The contemporary history of teaching Turkish as a foreign language (hereinafter TFL) dates back to the end of the 1970’s. The pioneers in this field are Hikmet Sebüktekin (“Turkish for foreigners: a linguistic approach”, 1969) and Kenan Akyüz (“Turkish for foreigners: speaking, reading”, 1979) from Bosphoros

University and Ankara University respectively (Erdem, 2009; Çiftçi and Demirci, 2018). Despite the fact that since then a rather large number of TFL textbooks and grammar books have appeared, practically no attempts have been made to create a Turkish learner’s dictionary. The literature review reveals nothing but a few theses buried in the bowels of the National Thesis Center of the Council of Higher Education of Turkey, namely “Glossary for Teaching Turkish as a Foreign Language” by Emrah Özcan (2006), “Preparing the basic vocabulary of Turkish for foreigners” by Ufuk Aşık (2007) and “Corpus based online learner dictionary: A headwords for adjectives” by Burak Tüfekçioğlu (2013). Whereas we strongly believe that the use of a learner’s dictionary can promote learner autonomy. By encouraging the intelligent and self-guided use of appropriate dictionaries, learners become more independent, which is one of the core goals of language educators.

Besides the above-mentioned theses there is a recently published “Beginner’s Dictionary of Turkish for Foreigners (A1)” by Engin Yılmaz. However, the examination of the dictionary reveals that it was prepared without taking into consideration the target users and the basic principles of learner lexicography. According to the preface, the database of the dictionary is made up of the words that appear in the textbooks published by Turkish learning centers of Yunus Emre Institute, Gazi University and Istanbul University. One of the main flaws of the dictionary is the complexity of the definitions which mostly results from the use of verbal adjectives and adverbs as well as voice affixes. For example, *açık büfe* (buffet) is explained in the following way: *Yiyecek ve içeceklerin serbestçe seçilip alınabildiği tarzda olan* (being the kind of that food and beverages can be freely chosen and bought). As we can see, the definition is overloaded by the grammatical constructions which are far beyond the understanding of A1 learners. The fact that *açık büfe* is labelled as an adjective is also quite arguable since it is a noun phrase and is often used as such, e.g. *Açık büfede, salata, soğuklar, zeytin yağlılar, çorba, iftar tabağı, ana yemek ve tatlı yer alıyor. / Hafta içi olmasına karşın içeride 5 - 6 masa dolmuş, yerimize geçtikten sonra açık büfeye yöneldik. / Basın mensupları maç boyunca açık büfeden istediğini ücretsiz alabiliyor.* The dictionary articles include abbreviations which demonstrate which of the textbooks the example sentence was borrowed from (e.g. *İÜ, YİTDK-A1,46*). However, this information is no use to the learners and does nothing but distract their attention. At last, the author of the dictionary has made an attempt to make it bilingualised by including the English translation of the headwords. However, in order to include foreign language equivalents, one has to have a good command of that language, otherwise this enterprise is going to end up with an inaccurate translation (e.g. open buffet instead of just buffet for *açık büfe*, big sister instead of elder sister for *abla*). Consequently, in spite of the author’s statement that “while preparing ‘Beginner’s Dictionary of Turkish for Foreigners (A1)’ the most important point was making this dictionary ‘user-friendly’, that is ‘learner-friendly’”, the dictionary turns out to hardly deserve this epithet. We believe that the discrepancy between the author’s intention and the final product stems primarily from the ignorance of the target users.

So who might the target users of the monolingual Turkish learner's dictionary be? Recently, the number of learners of Turkish language has grown due to the influx of Syrian refugees, as well as students benefiting from "Türkiye Scholarships", especially students from Africa and the Asia-Pacific region. These are the countries where turcological studies are either not strongly developed or don't exist at all. As a result, one is unlikely to find reliable bilingual dictionaries prepared for those coming from those countries. Therefore, our goal is to create a dictionary aimed at a wide range of learners of Turkish as a foreign language. We do keep in mind that the learners' approach to using a monolingual dictionary is influenced by the speakers' native language as well as the linguistic and lexicographical traditions and culture of their native country and that "the mother tongue should play a decisive role in the way the foreign language is presented ... for the target users" (Gouws, 2015: 349). However, at this stage this doesn't seem to be a realizable goal.

However, there is one major obstacle in the way of creating a monolingual Turkish learner's dictionary, namely the language itself. The fact that Turkish is a synthetic language possessing a lot of morphemes makes it hard for lexicographers to write an easy-to-understand definition. For example, the Oxford Advanced Learner's Dictionary defines the word "hope" as "a belief that something you want will happen", which is a rather clear and comprehensible definition. Yet if we decide to make the same definition in Turkish, we'll have to use at least two verbal adjectives which are learnt at level B2 (*İstediğiniz bir şeyin olacağına dair bir inanç*). Not using verbal adjectives and adverbs is also an option but it forces us to come up with several sentences instead of one which makes the definition sound a bit clumsy and unnatural: *Bir şey istiyorsunuz. Bu şey gerçekleşecek diye inanıyorsunuz. Bu umuttur.* (You want something. You believe that this will happen. This is hope).

Since 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" (Tarp, 2014: 244), the question then arises as to whether there is a need to create a monolingual Turkish learner's dictionary if a significant part of the target users is not capable to use it.

METHOD

In order to determine at what level learners of Turkish master the grammatical structures to be used in definitions, i.e. verbal adjectives and adverbs and voice affixes, four most commonly used TFL textbooks were studied. These are *Gazi TÖMER Yabancılar İçin Türkçe*, *Yeni Hitit Yabancılar İçin Türkçe*, *İstanbul Yabancılar İçin Türkçe*, and *Yedi İklim Türkçe*. It was found out that the grammatical competence required to process the definitions written in Turkish is supposed to be acquired at level B2 (Upper-Intermediate level). According to CEFR, learners at this level “can understand most short stories and popular novels” and might require a dictionary for specialised or unfamiliar texts (Council of Europe, 2009: 239). Therefore, in order to perform the experiment, an excerpt from the short story by Haldun Taner titled *Bir motorda dört kişi* (“Four people in a motorboat”) was chosen on the grounds that it contained words which were unknown and irretrievable from the context.

Then a pilot study was conducted with a 24-person participant group. All of the participants were Upper-Intermediate (B2) learners studying Turkish in SBÜ TÖMER. Countries of origin of the participants were as follows: Turkmenistan (7), Afghanistan (3), Burkina Faso (3), China/East Turkestan (2), Iran (2), Syria (2), Yemen (2), Ethiopia (1), Kazakhstan (1), and Nepal (1). The subjects were first asked to fill out a questionnaire. Keeping in mind the weak points of questionnaire-based research into dictionary use, namely the fact that questionnaire results “are often a measure of respondent’s perceptions, rather than objective fact” and that “researcher and respondent do not necessarily share the same terms of reference” (Nesi, 2000: 12), we only included the questions aimed at eliciting personal information about the respondents, i.e. their sex, age, education status, native country, mother tongue, language level and three factual questions concerning dictionary use: Do you use a dictionary? If so, which dictionary do you use? What is the format of the dictionary you use (printed/online/application)? After filling in the form, the subjects were asked to read the above-mentioned text and underline all the words whose meanings they didn’t know or couldn’t infer.

As a result of the piloting, seventeen target words were selected from those which had been most frequently underlined. Then the entries of the selected lexemes were prepared based on the data of the Turkish National Corpus and Sketch Engine. All senses of polysemous words relevant to the learners were given. In cases where the definitions were thought to be insufficient, some illustrations were also included in the corresponding articles. Then the experiment was conducted. The participant group comprised 31 Upper-Intermediate (B2) Turkish learners studying in İstanbul University Language Center. Subjects’ countries of origin were: Afghanistan (3), Egypt (3), Kazakhstan (3), Palestine (3), Iran (2), Iraq (2), Syria (2), Algeria (1), Bangladesh (1), Bosnia and Herzegovina (1), Ghana (1), Kosovo (1), Kyrgyzstan (1), Macedonia (1), Montenegro (1), Pakistan (1), Russia (1), Somalia (1), Spain (1), and Sri Lanka (1). After filling out the aforementioned questionnaire, the subjects were told to complete three tasks measuring their receptive and productive skills. They were encouraged to use the mini dictionary consisting of the articles earlier prepared

by the author. The tasks assessing the receptive skills of learners included matching the words with the pictures (Task 1) and finding sentences in which the underlined words were used in the same sense as in the given short story excerpt (Task 2). The final task required consulting the dictionary if necessary and writing eight sentences using the newly learned words (Task 3).

In Task 1 we intentionally included the pairs of pictures that could cause confusion among the subjects of the experiment either because of the similarity of the objects they depict and their definitions (e.g. a projector and a streetlamp, a sheep and a goat, linoleum and parquetry, a door knob and a door handle, to crouch and to sit cross-legged, a pipe and a cigarette, a deck and a cabin) or the possible meaning that could be inferred from the constituent parts of the words (the pictures of an anti-aircraft gun and a fighter aircraft for *uçaksavar* where *uçak* stands for a plane, a firefly and a sparkler for *ateşböceği* where *ateş* stands for fire).

In Task 2 subjects were given five polysemous words (*savurmak, taramak, yanık, yapışmak, yarmak*) used in four different meanings each and asked to identify the sentences containing words with similar meaning. All of the sentences represented authentic language and were borrowed from the Sketch Engine.

In Task 3 subjects were permitted to use any of the seventeen target words in any of the senses to make up eight sentences.

RESULTS

In response to the question, “Do you use a dictionary?”, all the respondents answered affirmatively. When asked to specify which dictionary they used, 15 respondents mentioned using *Google Translate*; 8 respondents, *Sesli Sözlük*; 6 respondents, *Tureng*; 5 respondents, printed bilingual dictionaries; 1 respondent, *TDK Sözlük*.

The experiment aimed at finding out whether the use of a monolingual Turkish learner’s dictionary contributed to Upper-Intermediate students’ comprehension of separate words and words used in context and their ability to produce new utterances revealed the following results:

1. In Task 1 none of the subjects could match correctly all of the given words with the corresponding pictures. The words which most of the subjects could successfully identify were *dazlak, ışıldak*, and *pipo* (Table 1).

Table 1

Percentage of successful completion of Task 1

Words	The percentage of correct matches
Ateşböceği	74%

Çımacı	74%
Çömelmek	35%
Dazlak	96%
Güverte	67%
Işıldak	93%
Kamara	64%
Karaman	70%
Muşamba	64%
Pipo	77%
Topuz	58%
Uçaksavar	48%

As expected, the most confusing words and pictures were those denoting an anti-aircraft gun and a fighter aircraft, crouching and sitting cross-legged, and a firefly and a sparkler (Table 2). Besides there were some completely irrelevant matches such as dockman – parquetry, deck – pipe, linoleum – deck etc.

Table 2

The most confusing pairs of words and pictures

Word	Picture	Number of mismatches
Anti-aircraft gun	Fighter aircraft	16
Crouching	Sitting cross-legged	10
Firefly	Sparkler	5
Projector	Sparkler	3
Anti-aircraft gun	Door knob	3
Door knob	Door handle	3
Sheep	Goat	2
Linoleum	Parquetry	2
Crouching	Streetlamp	2
Dockman	Deck	2
Pipe	Cigarette	1
Deck	Parquetry	1
Cabin	Deck	1

2. In Task 2 the subjects also scored low in the ability to detect similar meanings of polysemous words used in context. The only word that overcame the 50-percent threshold was *yarmak* (Table 3).

Table 3

Percentage of successful completion of Task 2

Words	The percentage of correct matches
Savurmak	6%
Taramak	32%
Yanık	45%
Yapışmak	35%
Yarmak	54%

3. When evaluating Task 3, two criteria were used: vocabulary control and grammatical accuracy. According to CEFR, an Upper-Intermediate learner shows a relatively high degree of grammatical control and lexical accuracy, “though some confusion and incorrect word choice does occur without hindering communication” (Council of Europe, 2009: 123) . However, only eleven of the thirty one subjects could produce between 4 and 8 grammatically and lexically correct sentences. Fourteen subjects could write between 1 and 3 sentences. Five subjects failed to make up any sentences. At last, one of the subjects rewrote the examples sentences substituting one word for another.

Errors made by the subjects can be classified into 2 categories:

1. Grammatical and lexical errors that are beyond the scope of this study unless they concern the use of the target words.
2. Errors resulting from the misunderstanding of the target words, e.g. *Sandalın çımacı çözülmüş ve şimdi nereye gittiğini bilmiyoruz.* The dockman of the boat has come loose and now we don't know where it has gone. *Başarılı olduğundan yanık insan kokusunu hissediyor musun?* Do you feel the smell of burned people because you're successful?

DISCUSSION

The fact that the majority of respondents mentioned *Google Translator* when asked about which dictionary they used shows that they see no difference between a dictionary and a machine translation service and that they might not have any other lexicographic tools available. Besides that, about half of the respondents, none of whom are native English speakers, turned out to use English-Turkish online dictionaries or dictionary applications (*Sesli Sözlük, Tureng*). This means that they access the meaning of L3 words via the

previously learned L2, which makes Turkish language educators rely on their students' proficiency in English.

The mismatches in Task 1 may have resulted from the subjects' reluctance to consult the dictionary and their reliance on the resemblance between the object on the picture and the word. Another reason for subjects' failing to complete this task successfully might be not knowing the words constituting either the genus or the differentia of definitions. At last, the existence of completely irrelevant matches makes us doubt some of the subjects' conscientiousness and language proficiency.

The word which was identified correctly by 96% of the subjects was *dazlak* (baldhead). This could be due to the simplicity of the definition ("having no hair") and/or the absence of any confusing pictures.

Subjects' being unable to find the words used in the same meaning in Task 2 can be explained by several factors, e.g. an overall difficulty of the task for a non-native speaker, the existence of too many unfamiliar words in the sentences to be matched, the insufficiency of definitions and example sentences, the discrepancy between the subjects' declared lexical competence (B2: Has a good range of vocabulary for matters connected to his/her field and most general topics) and their actual vocabulary knowledge, and the subjects' unwillingness to fulfil this task.

Task 3, which engaged learners' productive skills, revealed that some of the definitions must have seemed unclear to the subjects, thus leading to miswording. Besides, not all the subjects turned out to have a satisfactory command of written Turkish as well as the expected degree of grammatical control and lexical accuracy, which made us think about whether the results of our experiment can be regarded as an indicator of the efficiency of using a monolingual Turkish learner's dictionary. The students who took part in the experiment had received 448 hours of instruction to enter the B2 CEFR level. However, according to Pearson, for example, fast students need at least 760 hours to reach the B2 level, whereas slow students acquire the same level of language proficiency in 1996 hours. Therefore we can conclude that some of the subjects of our experiment were at level B2 only on paper, and they shouldn't be considered as the target users of a monolingual Turkish learner's dictionary.

CONCLUSION

The present study was an attempt to validate the effectiveness of using a monolingual Turkish learner's dictionary and the necessity to create one. It was suggested that a monolingual Turkish learner's dictionary is more suitable for students at level B2, since grammatical structures most often found in definitions are studied at this level. In accordance with the reading competencies described in CEFR, an excerpt from the short story by Haldun Taner was chosen and a pilot study was conducted in order to determine the target words to be included in the main study. The entries of the selected lexemes were prepared based on the data of the Turkish National Corpus and the Sketch Engine. The study participants had to fill out a questionnaire and to complete three tasks measuring their receptive and productive skills.

It was found out that many TFL students rely on Google Translator to perform both encoding and decoding tasks. However, despite the fact that Google Translator ensures fast access to translation and meanings, this doesn't imply that the translation and meanings it provides are always accurate and relevant. Besides Google Translator, the subjects turned out to use online English-Turkish dictionaries or dictionary applications, though none of them was a native-language English speaker. This suggests that the participants of the study are either unaware of the existence of bilingual dictionaries associating words in Turkish with the words in their native language or reluctant to use them. This also implies that there might not be any satisfying electronic bilingual Turkish dictionaries. It also should be mentioned that students being forced to reach the meaning of L3 through L2 binds their success in learning Turkish with their proficiency in English. Hence TFL students should be reminded of the limitations of the tools they use in language learning, and more efforts should be made to create up-to-date bilingual and/or bilingualized Turkish dictionaries.

As for the experiment, the subjects proved not to have benefitted from the use of the mini dictionary we prepared. They showed low performance in both receptive and productive activities, and some of the subjects completely failed to fulfil the given tasks.

Nonetheless, these results must be interpreted with caution and a number of limitations should be borne in mind, the first one being the size of the sample and the second one being the subjects' linguistic competence. We believe that at this stage the results of the experiment should not be considered indicative of either the efficiency or inefficiency of using a monolingual Turkish learner's dictionary and that it should be repeated using a new, larger group of subjects. We assume that the target users of a monolingual Turkish learner's dictionary are Upper-Intermediate and Advanced learners provided with sufficient instruction and enough time to acquire the language. Therefore, in subsequent experiments, a strict selection of participants based on the language test results is viewed as necessary.

References

Aşık, U. (2007). Yabancılar için temel Türkçe sözcük varlığının oluşturulması (Master's thesis, Dokuz Eylül University). Available from Tez Bankası. (No 211630)

Benigno, V., De Jong, J.H., Van Moere, A. (2017). How long does it take to learn a language? Insights from research on language learning. Retrieved from Researchgat

https://www.researchgate.net/publication/318233428_How_long_does_it_take_to_learn_a_language_Insights_from_research_on_language_learning

Council of Europe (2009). Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR). Strasbourg, France: Council of Europe. Retrieved from <https://rm.coe.int/16802fc1bf>.

Çiftçi, Ö., Demirci, R. (2018). Türkçenin Yabancı Dil Öğretimiyle İlgili Bir Kaynakça Denemesi. *Turkish Studies*, 13 (28), 265-339.

Erdem, İ. (2009). Yabancılarla Türkçe Öğretimiyle İlgili bir Kaynakça Denemesi. *Turkish Studies*, 4 (3), 888-937.

Gouws, R. (2015). Who are the target users of monolingual learner's dictionaries? *Tydskrif vir Geesteswetenskappe*, 55 (3), 343-355. doi.10.17159/2224-7912/2015/v55n3a2

Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer Verlag.

Özcan, E. (2006). Başlangıç düzeyi yabancı dil olarak Türkçe öğretimi için sözlükçe çalışması (Master's thesis, Yıldız Teknik University). Available from Tez Bankası. (No 188548)

Tarp, T. (2014). Dictionaries in the Internet Era: Innovation or Business as Usual? *Alicante Journal of English Studies*, 27, 233-261.

Tüfekçi, B. (2013). Derlem tabanlı çevrim içi Türkçe öğrenci sözlüğü: Önadlar A madde başı (Master's thesis, Mersin University). Available from Tez Bankası. (No 345244)

ENRICHING SYNSETS IN TAMIL WORDNET: PARADIGM SHIFTS IN LEXICOGRAPHY

S.Arulmozi

University of Hyderabad

Abstract

Lexicography, being sui generis, is informed by various disciplines like information science, literature, publishing, philosophy, cognitive science, and historical, comparative and applied linguistics. Since dictionaries constitute the ontological core of lexicography, lexicological studies reveal the epistemological organizations, its changes and reasons for such epistemological shifts. The Indian tradition of lexicography was philosophical in nature which emphasized both on philosophy of language and grammar as it is exemplified in *nikantu* or *nirukta*. Among others, one of the claims this paper would argue for is that the missionary production of dictionaries (monolingual/bilingual/trilingual) in India is one of the major paradigm shifts that happened in areas of both philosophy of language and lexicographic tradition — a paradigm shift that witnessed a transition from the philosophical understanding of communication embedded in the lexical tradition to the utilitarian scope of administrative communication. The format of presentation, ordering of lexical entries, and the composition of dictionaries was modeled after westernized lexicographical tradition. This paper would further argue for the claim that lexicographic work in India finds itself yet in another epistemological shift that starts with electronic computation. Researches on building electronic dictionaries with the help of a representative corpus are located just at this niche. Though there are major researches happening in major Indian languages for the creation of various corpora, the present paper claims that the spinoff of such efforts such as electronic dictionaries, lexical resources, WordNets, are not exhaustive in nature unlike traditional dictionaries and lexicons. The paper claims that the traditional skillset, based on the philosophical understanding of imbibing the worldview into the word, is missing in the product-driven principle of electronic computation. These claims will be exemplified by showcasing how the existing Tamil WordNet can be enriched by drawing examples from print and corpus-based Tamil dictionaries.

Key Words: Tamil, WordNet, Synset, Nighantu, Corpus.

1. Introduction

The Princeton English WordNet (Fellbaum, 1998) is one of the most resourceful semantic lexical database in English. Its main advantage is that it is hand-crafted, so data stored within its semantic network are of high quality. It is widely used as a resource in many NLP applications such as Information Retrieval, Word Sense Disambiguation, etc. The continuous expansion of the multilingual information society with a growing number of new languages present on the World Wide Web has led in recent years to a pressing demand for multilingual applications. To support such applications, multilingual language resources are needed, which however require a lot of human effort to be built. For this reason, the development of language independent resources which factorize what is common to many languages, and are possibly linked to the language-specific resources, could bring great advantages to the development of the multilingual resources in Indian languages.

Princeton's English WordNet inspired extensive development of WordNets in European languages, EuroWordNet (Vossen, 1998) and also in other languages across the globe including WordNets in Indian languages, IndoWordNet (Pushpak, 2010).

In this paper, a brief account on the paradigm shifts in Tamil lexicography works is pinpointed. Although Tamil WordNet was built using expansion approach (i.e. providing equivalent synonym sets from Hindi to Tamil), the equivalent senses are very limited thus the meaning of the concepts from in Tamil WordNet are scarce. In this paper, it is claimed that utilizing the traditional print and corpus-based online dictionaries will help in enriching the synonym sets in Tamil WordNet. The paper is organized as follows: Section 2 details about the paradigm shifts in Tamil lexicography. Section 3 deals with a discussion on the construction of IndoWordNet in general and pinpoints a few problems faced during the construction of synsets in Tamil. Section 4 briefly lists the missing links in Tamil WordNet followed by the ways in which Tamil WordNet synsets can be enriched utilizing the earlier Tamil lexicographical works. The last section summarizes the work.

2. Tamil Lexicography & Paradigm Shifts

Periodic revolutions are generally called as "Paradigm shifts" (Kuhn, 1970) and this is true in case of Tamil which has undergone turns, particularly Tamil Lexicography. *tolkaappiyam* can be regarded as the first authoritative work which contributed and paved the way for the Tamil Lexicography. *tolkaappiyam* is considered as the first available old Tamil master piece which consists of pages that deal with the meanings of words. The chapter on *ezhuttatikaaram* deals with the orthography, *collatikaaram* deals with the etymology of words and *porulatikaaram* deals with prosody, rhetoric and sociology. *uriyiyal* which is part of *collatikaaram* deals particularly with the words; but meanings which are given for words does not exhibit the alphabetic arrangement of words.

The first paradigm shift in Tamil lexicography can be attributed to the *nikantu*, viz. *ceentan tivaakaram*, a metrical dictionary in twelve parts; *pinkalanikantu*, in ten parts ascribed to the son of *tivaakaran*;

cuutaamani, as in *tolkaappiyam*, here too words are not arranged in alphabetic order. Nikantu arrangement is usually in poetic form, the only exception is *cuudamani nikantu* which classifies words according to rhymes (*ethukai*). In these words were classified into 18 kinds such as *kakara ethukai* to *nakara ethukai*. Alphabetic ordering is found only in case of the first or second letter of words. In *akarathi nikantu* words were arranged in alphabetic order and hence the name. This is the first lexicon of Tamil. One can see that only difficult or hard words found place in *nikantu*.

The next shift comes in the contribution by the missionary works. Tamil-Portuguese lexicon (Antem de Proenca); Cadura akaraathi (Father Beschi, popularly known as veeramaa munivar) – quadruple lexicon. This is a pioneering work for the dictionaries that gave meanings to Tamil words in Tamil. It is divided into four parts i) peyar (homonyms), ii) porul (synonyms), iii) thokai (nominal groups) and iv) thodai (rhymes). Then came the works of Fabricius's Tamil-English dictionary, Tamil-English dictionary by Rottler, English-Tamil dictionary by Winslow. These three dictionaries gave meanings in English for Tamil words and in Tamil for English words. In all these dictionaries, most of the commonly used words found place.

The third paradigm shift is seen in the growth of vocabulary lists in the form of Tamil monolingual and bilingual dictionaries, Dictionary of technical terms and Glossaries. The next shift could be attributed to the monumental work on Tamil lexicon published by Madras University. The Tamil lexicon (first edition, 1933) had approximately 1.25 lac words and is said to be the most comprehensive dictionary of the Tamil language. For a detailed history of Tamil lexicography, see, Vaiyapuri Pillai (1933). Due to the information technology developments, Tamil lexicography has crossed its traditional methods of compiling dictionaries to using corpus-based dictionaries. Mention should be made about CreA's Modern Tamil dictionary, Pals dictionary and Tamil WordNet.

3. Tamil WordNet

In the present situation, WordNet is taken over all the earlier practices of dictionary making and replaced by the technological developments, especially WordNet with all its implications. Tamil WordNet is an attempt to build a lexical network for the Tamil language along the lines of the English WordNet/Hindi WordNet so that it can be used as a tool for enhancing the performance of MT systems involving Tamil. Each word will be assigned a set of all possible senses it can take. It will also capture various relationships between the words by networking the sense of these words in an appropriate manner using the relationship as a function. These word-level relations include synonymy, antonymy, hypernymy, hyponymy, meronymy and holonymy. A Machine Translation system having the source language as Tamil can effectively exploit these relations to resolve ambiguities in the text.

Tamil followed two different approaches in the construction of WordNets. In the first approach, it relied on Rajendran's (2001) Modern Tamil Thesaurus, which is based on Nida's (1975) Componential Analysis of Meaning. This work which is also available in the electronic form represents the ontological structure of

Tamil vocabulary. Tamil vocabulary is classified into four major domains: entities, abstracts, events and relationals based on the part-of-speech categories. Along the lines of Nouns in English WordNet, Tamil nouns are divided into several hierarchies, each representing a unique beginner. These multiple hierarchies correspond to relatively distinct semantic fields, each with its own vocabulary. Unique beginner corresponds roughly to a primitive semantic component in a compositional theory of lexical semantics. In Tamil WordNet, nouns are classified mainly into two: *parumaipeyarkaL* ‘concrete nouns’ and *aruvappeyarkaL* ‘abstract nouns’.

The second approach needs special mention as the construction of WordNet took a different turn as work on other Indian languages WordNets was happening simultaneously. WordNets in Indian languages are constructed using the expansion approach including Tamil. i.e. Hindi WordNet synsets are taken as a starting point of departure. The concepts provided along with the Hindi synsets are first conceived and appropriate concepts in Tamil are manually provided by language experts. The Tamil synsets are then built based on the concepts created keeping in view the three principles, viz. Minimality, Coverage and Replaceability. At the outset, this approach looks trivial and economical considering the interlinking of synsets of different languages. Since most of the synsets are translated from Hindi into Tamil, lexicographers have assigned only equivalents that map to Hindi synsets. For example, in Tamil WordNet, *aaTu* has 8 senses (3 noun senses and 5 verb senses). But if you look at the traditional dictionaries, particularly the comprehensive etymology of the Tamil language, one can see a whole list of senses with all extended meanings covered. A lexical database like Tamil WordNet should take cues from such elaborative works for enriching the lexical resources.

For instance, English WordNet lists 35 senses for the word *go* which includes 4 nominal senses, 30 verbal senses and one adjectival sense. Tamil, a South-Dravidian language lists 9 verbal senses for the word *poo* ‘go’ (as compared to Hindi WordNet which lists 2 nominal senses, 16 verbal senses and 2 adjectival senses for the word *chalna* ‘go’). There is no guarantee that only these are the possible senses for the word under consideration. As we know language is dynamic and not static. So there is always a possibility of expansion of the meaning of a word (i.e. addition of new senses) as the word may be used in new contexts.

In what follows, we present few examples from Tamil WordNet, Comprehensive etymological dictionary and Corpus-based online dictionary (CreA) for demonstrating the senses across different lexicographical works.

4. Enriching Tamil WordNet

Tamil WordNet as mentioned above used expansion approach (in the second approach) wherein synsets from Hindi WordNet and replaced by Tamil synsets. In most of the cases, very few senses (denoting the concept in Hindi and its translation in Tamil) are given. Tamil lexicographers mostly translated the Hindi

synsets using Hindi-Tamil bilingual dictionaries. Let us illustrate this point by drawing senses from Tamil WordNet, creA and etymological dictionary (peerakaramutali).

Illustration 1: pati (has 12 senses in Tamil WordNet)

Tamil WordNet	creA	peerakaramutali
<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. interpret something that is written or printed. 2. settle into a position, usually on a surface or ground. 3. gain knowledge or skills. 4. learn by reading books. 5. look at, interpret or say out loud something that is written or printed. 6. happen, occur, take place. 7. produce or leave stains. 8. follow a course of study 9. see. 	<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. (of snow, moisture) be covered with; (of dust, etc.) settle; form; gather. 2. be straightened; be firmly pressed. 3. be ingrained. 4. be submissive. 5. (of price, bargain) be settled. 	<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. to settle, as dust or sediment 2. to gather as cream 3. to rest, as clouds upon a mountain, to alight to roost, as birds. 4. to be subjugated; to be trained, disciplined or tamed. 5. to become orderly, settled, as handwriting. 6. to obey. 7. to bathe, to sink in water, to be immersed. 8. to close, as eyes. 9. to become compressed, flattened, as olas, leaves, leather. 10. to subside, as water. 11. to be joined, united. 12. to fall prostrate.
<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. an amount allowed or granted. 2. weight for scales. 	<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. step; staircase/(in a ladder) rung 2. stage 3. a cut (above) 4. a measure of 8 5. allowance paid to an employee in addition to the basic pay 	<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. step, stair; rung of a ladder. 2. grade, rank, class, order. sphere. 3. nature 4. stirrup 5. weight for scales. 6. a weight = 100 palam. 7. the ordinary measure of capacity = 8 ollocks, kottu of jaffna.

	<p>6. copy (of a book, document, etc.)</p>	<p>8. fixed daily allowance for food. 9. device, means. 10. state, condition. 11. manner, mode. 12. still or lintel. 13. body. 14. family, lineage. 15. fitness. 16. order. 17. low platform for conducting ceremonies. 18. reservoir of water.</p>
--	--	---

Illustration 2: aaTu (has 12 senses in Tamil WordNet)

Tamil WordNet	creA	peerakaramutali
<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. to extend, wave or float outward, as if in the wind 2. move or sway in a rising and falling or wavelike pattern 3. Move back and forth or sideways 4. Move in a pattern; usually to musical accompaniment; do or perform dance 	<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. (of sth. which is hanging or standing in a position) move in a swaying motion; move to and fro. 2. (of body) shiver, tremble. 3. vibrate 4. shake 5. (of a swing or one in a swing) go forward and backward; swing 6. perform (a dance, drama). 7. dance or move as if dancing 8. play (a game) 9. dance to the tune of 10. behave without restraint; have an intemperate life-style 	<p><i>Verb Senses</i></p> <ol style="list-style-type: none"> 1. to move, to wave, to swing, to shake, to vibrate. 2. to dance, to gesticulate. 3. to play 4. to bathe, to play in water 5. to go, to proceed 6. to practice, to preserve 7. to be born 8. to wander 9. to rotate 10. to become weak 11. to throb 12. to quiver 13. to shudder, to tremble 14. to be crushed in a mill 15. to fall 16. to cohabit 17. to be proud
<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. male goat 	<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. goat, sheep 	<p><i>Noun Senses</i></p> <ol style="list-style-type: none"> 1. goat 2. sharpness

As illustrated in the examples above, one can visualize that not all the senses that are available in Tamil vocabulary (as exemplified in the select dictionaries above) are enlisted in Tamil WordNet. It is clear that only very few senses are attested in Tamil WordNet, especially translation of equivalents from Hindi into Tamil. Senses which are available in traditional dictionaries and other online dictionaries are not taken into

consideration in Tamil WordNet. Many efforts have to be made to enrich Tamil WordNet to make it a valuable lexical resource that will be useful in NLP applications.

6. Conclusion

In the present work, we emphasized two important works viz. Comprehensive etymological dictionary of Tamil language and creA, corpus-based online dictionary that need to be consulted for enriching Tamil WordNet. We illustrated the differences between the existing Tamil WordNet alongside the two dictionaries as mentioned above. To prove our point, if enrichment of Tamil WordNet happens, it will populate the database of Tamil. From our discussion it is also clear from the consultation of dictionaries, Tamil exhibits polysemy and hence has more extended senses and these needs to be included in Tamil WordNet.

References

- Apresjan, J. D. 1973. *Regular Polysemy*. Mouton, The Hague.
- Bhattacharya, Pushpak. 2010. IndoWordNet. In: *Proceedings of the Seventh International Conference on Lexical Resources and Evaluation (LREC'10)*. ELRA Publication.
- CreA. 1998/2008. *taRkaalat tamizh akarati*. Chennai: CreA Publication.
- Cuyper, I & G.Adraens. 1997. *Periscope: the EWN Viewer*. EuroWordNet Project LE4003. Deliverable D008d012. Amsterdam: University of Amsterdam.
- Fabricius, J. P. 1972. *Tamil and English Dictionary*. Tranquebar: Evangelical Lutheran Mission Publication House.
- Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.
- Miller G.A. , R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography*, Vol 3, No.4, 235-244.
- Pustejovsky J. 1995. *The Generative Lexicon*. Cambridge, MA.The Hague: Mouton.
- Rajendran, S. 2001. *Modern Tamil Thesaurus (in Tamil)*. Thanjavur: Tamil University Publication.
- Santhanam, K. (ed.). 2007. *A Comprehensive Etymological Dictionary of Tamil Language (centamilc coRpiRappiyal peerakaramutali)*. Chennai: Govt. of Tamil Nadu Publication.
- Shanmugam Pillai, M. 1985. *tamil-tamil akaramutali. (in Tamil)*. Chennai: Tamil Nadu Text Book Society.
- Thaninayangam, Xavier S. (ed). 1966. *Antao de Proenca's Tamil-Portuguese Dictionary*. Kuala Lumpur: University of Malaya.

Vaiyapuri Pillai, S. (ed.). 1931. *arumporuL viLakka nikantu (in Tamil)*. Madurai: Tamil Sangam Publication.

_____. (ed). 1933. *Tamil Lexicon*. 6 Volumes. Madras: Madras University Press.

Vossen P. (eds.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Winslow, M. 1892. *A Comprehensive Tamil and English Dictionary of High and Low Tamil*. Madras: P.R. Hunt.

DICTIONARY OF ANCIENT TURKIC AND MONUMENTS OF MONGOLIAN RUNIC INSCRIPTION

Azzaya Badam

Department of Asian Studies, School of Arts and Sciences, National University of Mongolia

Otgonsuren Tseden

**Department of British and American Studies, School of Arts and Sciences, National University of
Mongolia**

Abstract

Mongolia is rich in culturally and historically precious stone monuments in connection with the Ancient Turkic which established its State in VI-IX AD and remained its name in the world history. Of them, the monuments of runic inscription have stood at a special position as a scientifically valuable proof of the Ancient Turkic history, language and culture. We, in Mongolia, have still discovered some new findings of the inscriptions that have reached 160 and more in number since the first discovery in 1889. For these years, we have used a couple of dictionaries: “Drevnetyurkski slovari (Древнетюркский словарь)” compiled by Russian scholars (Наделяев В.М., Насилов Д.М., Тенишев Э.Р., & Щербак А.М., 1969) in 1969 and “An Etymological Dictionary of Pre-Thirteenth-Century Turkish”, by Sir Gerard Clauson in 1972.

Since the time when the aforementioned dictionaries compiled, more than 90 runic inscriptions have been newly found in the land of Mongolia every year, which incorporate numbers of words and expressions we have no longer met before.

As a result, we should conduct some lexicological studies based on the relevant monuments of runic inscriptions and compile a dictionary of Mongolian runic inscriptions then, which brings some significance for the development of ancient Turkic vocabulary, the definition for unknown words and expressions, as well as some comparative studies on the Altaic languages in the sphere of semantics.

Keywords: Mongolia, Runic Inscription, Ancient Turkic, Vocabulary

I. Introduction

The runic inscription study in Mongolia occupies a special and important position in the Turkic studies in the world and its new findings, discovered in the territory of Mongolia every year, are a kind of source materials which are eagerly awaited by the scholars and researchers who conduct their research in this field.

All the runic inscription monuments newly found in Mongolia have to be registered for their valuable fact. In fact, every newly found monument adds numbers of new words and expressions to the ancient Turkic language fact and vocabulary, but all of them have not been registered and introduced yet. The first registration was published in 2010 (Azzaya Badam, 2010; Баттулга Ц. & Аззая Б., 2009). Within these nine years, that is, between the years of 2010 and 2018, many monuments which are included in the category of small/minor monument of the runic inscription have been newly found, and deciphered by our scholars and researchers. However, the lexicographical research for these newly found monuments can be said to be comparatively limited to only just list of the new words noted in the relevant monuments.

As a result of them, this paper focuses on a description and a lexicographical study of the newly registered runic inscription monuments in Mongolian.

Within the years, totally 29 runic inscription monuments were newly found in 22 different places of Mongolia, which have made their number reach 166. There is a list of the runic inscription monuments newly found in 2010-2018:

II. Method

For the study of the runic inscription monuments date back to the periods of the Ancient Turkic and Uighur, there has been no any other sources excluding “Drevnetyurkski slovari (Древнетюркский словарь)” compiled by Russian scholars (Наделяев В.М., Насилов Д.М., Тенишев Э.Р., & Щербак А.М., 1969) in 1969 and “An Etymological Dictionary of Pre-Thirteenth-Century Turkish”, by Sir Gerard Clauson in 1972. But, since the time when they both were printed, more than 100 monuments have been found in the territory of Mongolia today.

In the framework of this study, we are planning to register all the runic inscription monuments found in 2010-2018, on the basis of the previous research work and reports as well as other oral sources.

In order to observe the description, content and lexicological properties of the monuments, all of them will be classified as the classic and small of the Mongolian runic inscription, their survey and lexicological elements will be done.

a. The classic monument of Mongolian runic inscription (2008-2018)

In the category ‘Classic’, we have chosen the monuments which are considered to be perfect in their making, rich in their vocabulary and complete in their content, like the following two monuments:

1. *Three monuments of Dongoyn Shiree Complex:*

The monuments of Dongoyn Shiree Complex are located near Delgerkhaan mountain, in Tuvshinshiree soum, Sukhbaatar province. They were firstly found by a surveying expedition led by D.Tumen, a professor of the Department of Archaeology and Anthropology, National University of Mongolia in 2010 and proved to be an ancient Turkic inscriptions on the monuments, by an exploring expedition of which members were Ts.Bolorbat (PhD), a research worker of the Institute of History and Archaeology, Mongolian Academy of Sciences and J.Gerelbadrakh (PhD, Prof), the head of the department of History, Mongolian State University of Pedagogy, then.

These three monuments of the Dongoyn Shiree Complex have kept 72 lines of inscription in total, with 3857 characters, 1193 single words and 86 tamgas on their bodies (МӨНХТУЛГА Р., 2018, р. 69; МӨНХТУЛГА Р. & Оосава Т., 2015, р. 53)

2. *Monument of Khirgisiyn Ovoo:*

Our research team of which members are T.Iderkhangai (Department of History and Anthropology, Ulaanbaatar University) and Ts.Battulga (Department of Asian Studies, National University of Mongolia) first discovered a fragment of the monument with a runic inscription in a place, on the northwest bench of the Khukh Erigiin Garam (Khukh Erig Ford) of the Baidragiin Gol (Baidrag River), at the distance of approximately 45-50 km to the northwest of Buutsagaan soum, Bayankhongor province - in May, 2016. The fragment has kept four wild goats (one small and three big) sculpted on its left side and eight lines of runic inscription (two, on the right and three, on each of back and front) on the other sides (Баттулга Ц. & Идэрхангай Т., 2016; Идэрхангай Т. et al., 2017; 마트돌가 체., 2017). Then professor Ts.Battulga deciphered some words and expressions and published with their interpretation and glossary in 2017 (Баттулга Ц., 2017b).

b. The small runic inscription:

In this category 'Small', the inscriptions that consisted of one or little more lines and unclear forms of characters, engraved on the surface of the rock. There are 25 inscriptions which are included in this category:

1. *Inscription of Baga Khayrkhan II /Khoыр Khavchig Uul II/* (found in Ikh Tamis soum, Arkhangai province, in 2010)
2. *Inscription of Taryat Winter-camp* (found on the rock of Taryat Winter-camp, Otsog Khedree mountain, Mankhan soum, Khovd province by Kh.Byambasuren, a curator of the Local Museum of Khovd province, and Ch.Enkhtor, a teacher of history, the secondary school of Mankhan soum in 2010 and deciphered by Ts.Battulga (Баттулга Ц., Бямбасүрэн Х., & Энхтөр Ч., 2010a, b))
3. *Inscription of Silver vessel* (deciphered by R.Munkhtulga and proved that it must have been belonged to someone before it was taken to the Government as the State property in 2010 (Munkhtulga R., 2013, pp. 27–28); (Мөнхтулга Р., 2018, р. 175); (Mönhtulga R., 2016; Munkhtulga R., 2013))
4. *Inscription of Jirimiyn Hudag II* (Two lines of runic inscription engraved on the rock, the Del mountain, Ulziit soum, Dundgovi province, was deciphered by Ts.Battulga in 2010 (Баттулга Ц., 2010, pp. 140–141); and another line of inscription which is on the west of the main body of the inscription was deciphered by B.Azzaya in 2016 (Аззая Б., 2016, р. 578))

5. *Inscription of Khirgisiyn Hooloy* (found in the place called *Khirgisiyn Hooloy*, Khotont soum, Arkhangai province in 2011 and deciphered by A.Ochir and Ts.Battulga (Battulga Tsend, 2016; Баттулга Ц. & Очир А., 2011))
6. *Inscription of Dalt* (found in the place Dalt, Bombogor soum, Bayankhongor province, and deciphered by Ts.Battulga in 2012 (Эрдэнэ М. & Баттулга Ц., 2012))
7. *Inscription of Khotgor Khag* (found in the place Khotgor Khag, Erdeneburen soum, Khovd province, and deciphered by Ts.Battulga in 2013 (Баттулга Ц., 2016b))
8. *Inscription of Del Uul V* (four runic inscriptions had been found in Del mountain, Tagt bag, Olziit soum, Dundgovi province, then the fifth inscription was found in 2014 and deciphered by teachers and students of the Department of Turkish studies in the same year, (Аръяажав Б. & Аззаяа Б., 2014))
9. *Inscription of Tavit winter-camp* (found in the place Tavit winter-camp, Nariinteel soum, Ovorkhangai province by a research worker, J.Tsambagarav in 2015 and deciphered by Ts.Battulga (Баттулга Ц. & Цамбагарав Ж., 2015, pp. 43–44))
10. *Inscription of Khulsana Am* (found in Erdene soum, Govi-Altai province by B.Ariyajav in 2015, and deciphered and discussed by R.Munkhtulga in 2015 (Mönhtulga R. & Ariyajav B., 2016 ; Мөнхтулга Р., 2018, p. 92))
11. *Inscription of Ivdey Deer Stone* (found by an expedition of Mongol – America joint project ‘The North Mongol’ in 2015 (Бүрэнтөгс Г., 2017) and first deciphered by Munkhtulga (Мөнхтулга Р., 2018, pp. 111–112) (Munkhtulga R., 2017))
12. *Inscription of Davirt Buuts* (found by an artist and movie director, B.Bayar in July, 2016 and deciphered by Ts.Battulga (바트돌가 체., 2017, pp. 94, 111) (Баттулга Ц., 2016a))
13. *Inscription of Bugat* (found by an artist and movie director, B.Bayar in 2016 and deciphered by Ts.Battulga (Баттулга Ц., 2017a, p. 47))
- Inscription of Ulaanchuluut* (first found and registered by research workers Ya.Tserendagva (PhD), R.Turbat and S.Dalantai in 2016)
14. *Inscription of Ulaanchuluut I* (found in 2016, deciphered by B.Azzaya (Аззаяа Б., 2017) and noted that there were five characters by Yu.Boldbaatar, Ya.Tserendagva and R.Turbat (Болдбаатар Ю., Цэрэндагва Я., & Төрбат Р., 2017, p. 78))
15. *Inscription of Ulaanchuluut II* (deciphered by B.Azzaya (Azzaya Badam, 2018) and noted that there were characters [T, t, L, l, d, m, s] by Yu.Boldbaatar, Ya.Tserendagva and R.Turbat (Болдбаатар Ю. et al., 2017, p. 78))
16. *Inscription of Ulaanchuluut III* (found in 2016 and noted that there were characters [P, J, B, Q, m, r, l] by Yu.Boldbaatar, Ya.Tserendagva and R.Turbat (Болдбаатар Ю. et al., 2017, p. 77))
17. *Inscription of Ulaanchuluut IV* (four lines of this inscription was first numbered and deciphered by Yu.Boldbaatar (Болдбаатар Ю. et al., 2017, p. 75) and then by B.Azzaya (Аззаяа Б., 2017))
18. *Inscription of Ikh Nart* (found in Dalanjargalan soum, Dornogovi province by a research worker of the Institute of History and Archeology, Mongolian Academy of Sciences and his

colleagues in 2016 (Цэрэндагва Я. et al., 2016, pp. 15–16) and then by R.Munkhtulga in 2018 (Мөнхтулга Р., 2018, pp. 124–126))

19. *Inscription of Urtyn Gol* (found on the small rock called Khakheeliin khad, Nomgon soum, Omnogovi province and studied by Ts.Battulga (Баттулга Ц., 2013) and a two line of another new inscription was found and deciphered by Ts.Battulga along with an artist and movie director, B.Bayar (Battulga Tsend, 2018))
20. Inscription of Asgat (one line of this inscription was found in a place called Asgat, Galuut soum Bayankhongor province in 2018 by a Mongol – China joint field researching expedition (Battulga Tsend, 2018))
21. *Inscription of Deed Tsohiot* (one line of this inscription was found in a place called Tsohiot in the basin of Tsagaan Turuut river, Galuut soum Bayankhongor province in 2018 by a researcher G.Batbold who worked for a Mongol – China joint field researching expedition and deciphered by Ts.Battulga in 2018 (Баттулга Ц. et al., 2019))
22. *Inscription of Dund Tsohiot I* (this is rich in tamgas and it has 2 runic inscriptions. It was found by a researcher G.Batbold who worked for a Mongol – China joint field researching expedition in 2018 and deciphered by Ts.Battulga (Баттулга Ц. et al., 2019))
23. *Inscription of Dund Tsohiot II* (found by a researcher G.Batbold who worked for a Mongol – China joint field researching expedition in 2018 and deciphered by Ts.Battulga (Баттулга Ц. et al., 2019))
24. *Inscription of Khanangiyn Buuts* (found in a place called Khanangiyn Buuts, in the basin of Tsagaan Turuut river, Galuut soum Bayankhongor province in 2018 by G.Batbold who worked in a Mongol – China joint field researching expedition 2018 and deciphered by Ts.Battulga (Баттулга Ц. et al., 2019))
25. *Inscription of Tasarkhay Onts spring-camp* (found in a place called Tasarkhay Onts spring-camp, in the basin of Tsagaan Turuut river, Galuut soum Bayankhongor province in 2018 and deciphered by Ts.Battulga (Баттулга Ц. et al., 2019))

III. Result

In 2010-2018, totally 29 runic inscription monuments were found and they can be divided chronologically as the following:

2010: 4 monuments

2011: 4 monuments, 3 of which are included in the category of the Classical

2012: 1 monument

2013: 1 monument

2014: 1 monument

2015: 4 monuments, 1 of which is included in the category of the Classical

2016: 7 monuments

2017: 1 monument

2018: 6 monuments

This classification, as just above, clearly tells that the runic inscription monuments are found in the territory of Mongolia every year.

IV. Discussion

We attempted to classify all the runic inscription monuments found in 2010-2018 as *the classical* - 4 and *small* – 25 based on their contents, character fonts and makings.

As for their lexicological elements, there are words belonging to 4 different parts of speech: noun (71), verb (30), numeral (10) and others (23).

a. Nouns:

This can be sub-divided here into 2 main parts: names and substantives.

1. *The names*:

In this classification, there are 35 names including personal name (17), title (11), tribal name (4) and proper name (4):

- Personal names: *ančir* ᠠᠨᠴᠢᠷ *Anchir* (Personal name) Dalt.; *apa* -ᠠᠯᠠ *Apa* (personal name) Dund.Ts I (Tsokh.); *bilgä* ᠪᠢᠯᠭᠠ *Bilge* (unknown whether it is a personal name or a title) Hirg.o W1: *bilgä čurīmya; čašabačayī* ᠪᠢᠯᠭᠠ ᠴᠢᠷᠢᠮᠤᠶ᠋ᠠ ᠴᠢᠰᠢᠪᠠᠴᠠᠶᠢ *Chasha Bachagi* (personal name) Uln.ch II 3; *äbiz* (~biz) ᠠᠪᠢᠵ *Ebiz* (*Abiz*) (personal name) (~ we) Hirg.h 2; *ič* ᠶ *Ich* (personal name?) Del.u V1; *öz ič o᠓; künčir* ᠶᠤᠵᠢ ᠶᠢᠨ *Kunchir* (personal name) Dalt.; *öz* ᠤᠵ *Uz~Öz* (personal name?) Del.u V1: *öz ič o᠓; o᠓* ᠣᠮ *On* (personal name?) Del.u V1: *öz ič o᠓; qar* ᠬᠢᠷ *Khar* (personal name) Deed.ts (Tsokh.): *qar čur; qaya* ᠬᠢᠷᠠ *Khaya* (personal name?) Uln.ch III 1;

- Uln.ch IV 1: *qaya čur*; *qu* ↓ *Qu* (personal name?) Tvt.u 2: *qu yän si*; *si* 𐰽𐰺𐰍𐰏 *Si* (personal name) Tvt.u 1: *qu yän si*; *toliy* 𐰽𐰺𐰏𐰍 *Tulig* (personal name) Hirg.o N1: *toliy qana*; *uruṇu* 𐰽𐰺𐰏 *Urungu* (unknown whether it is a personal name or a title) (*lit.trans: flag*) Dvrt.b; *yän* 𐰽𐰺𐰏 *Yen* (personal name) Tvt.u 2: *qu yän si*; *yägän* 𐰽𐰺𐰏𐰍 *Yegen* (personal name) Hirg.o N2: *yägänčur*,
 - Titles: *baya* 𐰽𐰺𐰏 *Baya*(title) (*lit.trans: minor*) Hirg.o S3: *boyla baya tarqan*; *beg* 𐰽𐰺𐰏 *Beg* (title) (*lit.trans: lord*) *begim e* 𐰽𐰺𐰏 𐰽𐰺𐰏 *Don.sh*; *boyla* 𐰽𐰺𐰏 *Boyla* (title) Hirg.o S3: *boyla baya tarqan*; *čab(?)* 𐰽𐰺𐰏 merit, fame, reputation Uln.ch III 1; *čur* 𐰽𐰺𐰏 *Chur* (title) Hirg.o N2: *yägänčur*; Uln.ch IV 1; Uln.ch IV 3; Deed.ts (Tsokh.); *čurimya* 𐰽𐰺𐰏𐰍 *Hirg.o W1: bilgä čurimya*; *künči* 𐰽𐰺𐰏 *Kunchi* (title?, tribal name?) Dalt.; *qan* 𐰽𐰺𐰏 *Khan* (title) Hirg.o S1; *qana* 𐰽𐰺𐰏 Hirg.o N1: *toliy qana*; *tarqan* 𐰽𐰺𐰏 *Tarqan* (title) Hirg.o S3: *boyla baya tarqan*; *yaryan* 𐰽𐰺𐰏 *Yargan* (title) Uln.ch I,
 - Proper names: *atač* 𐰽𐰺𐰏 *Atach* (proper name?) (*lit.trans: Father*) Ivd.g 1: *uzun atač tejirim*; *ušya* 𐰽𐰺𐰏 *Ushga* (proper name) Don.sh IIAW2; Don.sh IIAW2; *uzun* *Nzv* *Uzun* (proper name?) (*lit.trans: long*) Ivd.g 1: *uzun atač tejirim*; *yašqan* 𐰽𐰺𐰏 *Yashqan* (proper name?) (*lit.trans: young Father*) Uln.ch III 1,
 - Tribal names: *az* *Az* (tribal name) Htgr.h; *oγuz* 𐰽𐰺𐰏 *Oγuz* (tribal name) Hirg.o N1; *toquz* 𐰽𐰺𐰏 *Toquz* (tribal name) (*lit.trans: nine*) Hirg.o N1: *toquz oγuz*; *türük* 𐰽𐰺𐰏 *Turkic* (tribal name) Hirg.o S2; Hirg.o S3.

Of them, the inscription of Tavit winter-camp is the most interesting in a personal name *qu yän si*, as it is deciphered to be a personal name of a foreign language.

2. The substantives:

This can be sub-classified in their lexicology and semantic field, as the following:

- The 6 headwords of substantives related to hunting: *ab* 𐰽𐰺𐰏 hunting Don.sh IIAW2; *al* 𐰽𐰺𐰏 lower part Jrm.h II; *alan* 𐰽𐰺𐰏 flat field Jrm.h II; *aṇ* 𐰽𐰺𐰏 hunting Jrm.h II: *aṇim a* 𐰽𐰺𐰏 Don.sh IBW5; *ay* 𐰽𐰺𐰏 snare Jrm.h II; *ärüli ilör* with sign Jrm.h II,
- The 6 headwords of substantives related to fight or battle: *alp* 𐰽𐰺𐰏 hero Dvrt.b; *bäk* 𐰽𐰺𐰏 strong, firm; *er* 𐰽𐰺𐰏 soldier Dalt.; Don.sh IIAW2; *kü* 𐰽𐰺𐰏 merit, fame, reputation Dalt.; *sü* 𐰽𐰺𐰏 army: *süsi* 𐰽𐰺𐰏 Ikh.n 1; *yay* 𐰽𐰺𐰏 arc, bow: *yayči* 𐰽𐰺𐰏 Dund.Ts I (Tsokh.),
- The 7 headwords of substantives related to piety and pray: *bäñü* 𐰽𐰺𐰏 eternal Ivd.g 2; *igä* Master (lord) Htgr.h; *qut* 𐰽𐰺𐰏 good fortune, belssings Dund.Ts II (Tsokh.); Uln.ch III 2; *qutluγ* 𐰽𐰺𐰏 to be blessed, to have a good fortune Dvrt.b; Uln.ch II 2; *teñiri iriht* Heaven: *teñirim miriht* Ivd.g 1: *uzun atač tejirim*; *teñri* 𐰽𐰺𐰏 Heaven: *teñrim e* 𐰽𐰺𐰏 Don.sh IIAW1,
- The 13 headwords of substantives related to politics and society: *bodun* 𐰽𐰺𐰏 people Hirg.o S2; *boduniγ* 𐰽𐰺𐰏 Hirg.o N1; *buluṇ* 𐰽𐰺𐰏 directions: *buluṇiγ* 𐰽𐰺𐰏 Hirg.o W2; *eb* 𐰽𐰺𐰏

home(house): *ebim e* ႳႷႸ Don.sh; *el* Ⴛ country Don.sh IbW5; *elig* ႻႻ Uln.ch IV 3; *elim e* ႳႷႸ Don.sh IIW1; *äl* Ⴛ country: *älin* ႻႻ Hirg.o S1; *äb* Ⴘ home(house) Bgt.; *il* ႻႻ country: *ilin* ႻႻ Hirg.o S1; *is* Ⴛ power, effort: *isig* ႻႻ Don.sh IIW1; *törü* ႻႻ state: *törüsün* ႻႻ Hirg.o S1; *yaqa* ႻႻ edge Don.sh IbW5; Don.sh IIAW2; Don.sh IIAW2; Don.sh IIAW2; Don.sh IIAW3; *yär* ႻႻ place, territory, land Urt.g 1; *yärdä* ႻႻ Uln.ch IV 1,

- The 5 headwords of substantives related to human relationship: *a* he/she/it: *anij* Htgr.h ; *apa* ႻႻ elder sister Bgt.; *bäj* ႻႻ I Uln.ch IV 3; *biz* (~*äbiz*) ႻႻ we (~*Ebiz* (*Abiz*) personal name) Hirg.h 2; *ečim* ႻႻ my Father Del.u V1.

b. Verbs:

There are 30 verbs in total in the runic inscription monuments in their lexicology and semantic field.

- The 25 headwords of verb related to politics, society, and fight or battle: *art-* ႻႻ to expand (the territory): *art{t}i* ႻႻ Urt.g 1; *artatip* ႻႻ to annihilate Hirg.o S1; *ašun-* ~*ašun-* ႻႻ to win: *ašundi~ašuntü* ႻႻ Hirg.o W3; *bar-* ႻႻ to reach, to go, to come to: *bariñ* ႻႻ Dund.Ts II (Tsokh.); Mng.a; *bašla-* ႻႻ 1. to lead, 2. to start: *bašlayu* ႻႻ Hirg.o S2; *ber-* ႻႻ make an effort for ..., to struggle, to give: *berdim e* ႻႻ Don.sh IIW1; *bit-* ႻႻ 1. to write, 2. to engrave, to carve: *bitdim* ႻႻ Del.u V1; Tvt.u 3; Uln.ch II 3; Uln.ch IV 2; *bitdim* ႻႻ Hirg.h 2; *et-* ႻႻ to organize (an army) Don.sh IIAW2; *ettirtim* ႻႻ Don.sh IIAW2; *äg-* ႻႻ to go back: *ägmädi* ႻႻ Hirg.o S3; *it-* ႻႻ to pacify: *itmis* ႻႻ Hirg.o W2; *käl-* ႻႻ to come, to arrive: *kälip* ႻႻ Tvt.u 1; *kälmis* ႻႻ Khan.b; *kältäci* Htgr.h; *kältimiz* ႻႻ Urt.g 2; *olur-* ႻႻ to promote, to appoint: *olurt{t}i* ႻႻ Dund.Ts I (Tsokh.); *ör-* ႻႻ to struggle: *örti* ႻႻ Hirg.o W3; *qazyan-* ႻႻ 1. to occupy (state or territory), 2. to possess (something or belonging): *qazyanü* ႻႻ Hirg.o N2; *qil-* ႻႻ to create, to make: *qilin* ႻႻ Jrm.h II; *qisal-* ႻႻ to become fewer, to decrease in number, to be reduced Ikh.n 1; *qisalmiš* ႻႻ Ikh.n 1; *qiš-* ႻႻ to set up, to make ... reach : *qišdi* ႻႻ Hirg.o S2; *tačiq-* =*tašiq-* ႻႻ 1. to struggle, 2. to invade: *tačiqdi=tašiqdi* ႻႻ Hirg.o N2; *täg-* ႻႻ to reach: *tägmis* ႻႻ Hirg.o W2; *tobira* ႻႻ Go around Huls.a; *toqüt-* ႻႻ 1. to build, 2. to set up: *toqütdim* ႻႻ Hirg.o W1; *tut-* ႻႻ 1. to rule, 2. to control: *tur{di} =tutdi* [ႻႻ] Hirg.o S1; *tutdi* ႻႻ Hirg.o N1; *yoq et-* ႻႻ

to sacrifice, to mourn: *yoq etmiş* 𐰽𐰺𐰍:𐰺𐰏 Don.sh IIaW2; *yor-* 𐰺𐰏 to go: *yortī* 𐰽𐰺𐰏𐰺 Don.sh Dvrt.b; Uln.ch I; Uln.ch II 1; *ärmis* 𐰽𐰺𐰏 *lit.trans: past form of verb 'to be'* Uln.ch IV 1,

- The 3 headwords of verb related to piety and pray: *ada-* 𐰽𐰺𐰏 to sacrifice, to offer Uln.ch III 2; *bol-* 𐰽𐰺𐰏 to be (happy) Dvrt.b; Uln.ch II 2; *öl-* 𐰽𐰺𐰏 to go to Heaven, to pass away Del.u V1,
- The 2 headwords of verb related to hunting: *abla-* 𐰽𐰺𐰏 to hunt, *ablama* 𐰽𐰺𐰏 Jrm.h II; *bayla-* 𐰽𐰺𐰏 to tie, to bound, *bayla* [𐰽𐰺𐰏] 𐰽𐰺𐰏 Jrm.h II.

c. Numerals:

- The 10 headwords of numerals: *bir rib* one Ivd.g 2; *äki* 𐰽𐰺𐰏 two Hirg.o S3; *äki yägirmi* (twelve); *tört* 𐰽𐰺𐰏 four Hirg.o W2; *tümän* 𐰽𐰺𐰏 ten thousand Hirg.o S2; *üç* 𐰽𐰺𐰏 three Hirg.o S1; *yägirmi* 𐰽𐰺𐰏 twenty Hirg.o S3; *äki yägirmi*(twelve); *yätmiş* 𐰽𐰺𐰏 seventy Hirg.o S2; *yegirmi* 𐰽𐰺𐰏 twenty: *yegirmikä* 𐰽𐰺𐰏 Hirg.h 2; *yiti* 𐰽𐰺𐰏 seven: *yitinč* 𐰽𐰺𐰏 Hirg.h 1,

d. Others:

There are 23 headwords which can be impossible to be accounted for the category “the noun”, in their semantic fields, such as *anči* 𐰽𐰺𐰏 such Dalt.; *anta* 𐰽𐰺𐰏 there (*lit.trans: since then*) Hirg.o S1; Hirg.o S2; *ay* 𐰽𐰺𐰏 month, moon Hirg.h 1; *käsrä* 𐰽𐰺𐰏 after Hirg.o S1; *qasi* 𐰽𐰺𐰏 fence Jrm.h II; *qay* 𐰽𐰺𐰏 rock Uln.ch IV 4; *qaya* 𐰽𐰺𐰏 rock Uln.ch IV 1; *qaya* 𐰽𐰺𐰏 rock Uln.ch IV 4; *qayaqa* 𐰽𐰺𐰏 Tvt.u 4; *qir* 𐰽𐰺𐰏 steppe: *qirim* 𐰽𐰺𐰏 Asgt. ; *ud* 𐰽𐰺𐰏 Ox Del.u V1; *uluy* 𐰽𐰺𐰏 big, large: *uluyi* 𐰽𐰺𐰏 Uln.ch IV 1; Uln.ch IV 4; *uluyqa* 𐰽𐰺𐰏 to Ulug (proper name) Don.sh IIaW3; *uq* 𐰽𐰺𐰏 origin, ancestor Mng.a; *üze* 𐰽𐰺𐰏 upper Don.sh IIW1; *yalim* 𐰽𐰺𐰏 smooth (*smooth surface of vertical rock*) Tvt.u 4; Uln.ch IV 1; Uln.ch IV 4; *yaš* 𐰽𐰺𐰏 life, age: *yašča* 𐰽𐰺𐰏 Hirg.o W3; *yayiz* 𐰽𐰺𐰏 uneven surface of land Don.sh IIW1; *yer* 𐰽𐰺𐰏 land, homeland: *yerim e* 𐰽𐰺𐰏 Don.sh; *yičä* 𐰽𐰺𐰏 again Hirg.o S1; *yil* 𐰽𐰺𐰏 Year: *yilqa* 𐰽𐰺𐰏 Del.u V1; Hirg.h 1; *yiš* 𐰽𐰺𐰏 mountain range: *yišqa* 𐰽𐰺𐰏 Don.sh IIaW2; *yont(yunt)* 𐰽𐰺𐰏 Horse (*year*) Hirg.h 1; *yori* 𐰽𐰺𐰏 journey, travel: *yoriγ* 𐰽𐰺𐰏 Dund.Ts II (Tsokh.).

Apart from the lexicological properties and semantic fields of the words in the monuments of Mongolian runic inscription as above mentioned, there is another special thing which can no longer be neglected here, - that is the frequency of merely three words in the monuments of

Dongoyñ Shiree III - a classical monument: “𐰇𐰏𐰣 ~ bma ~ ebim e” (“Oh, my house!” (occurred 950 times)), “𐰇𐰏𐰤𐰣 ~ bgma ~ begim e” (“Oh, my lord!” (180)), and “𐰇𐰏𐰢𐰣 ~ yrma ~ yerim e:” (“Oh, my country!” (38)). The frequency of the words like just above mentioned reveals either good wishes for something or the ornament of the monument with some small stones.

... Three steles had 3857 characters, 1193 words and 86 *tamgas*. The inscriptions are written in Ancient Turkic language. Such as expressions as “: ebim e:” (“Oh, my house!”), “: begim e:” (“Oh, my lord!”), and “: yerim e:” (“Oh, my country!”) had been repeatedly on the surfaces of the steles in memory of the deceased person. The above praying words in the epitaph look like ornaments to pattern the giant steles (МӨНХТУЛГА Р., 2018, p. 179)

The researchers have proven that the three monuments have a runic inscription, but are studying the possibility of any inscription on other monuments. According to the previous research, there are 1193 words frequently occurred: 𐰇𐰏𐰢𐰣 ~ lɯma ~ el aɯim a ~ Oh, Dear State Hunting, 𐰇𐰏𐰤 ~ YQa ~ yaqa ~ edge on the 5th line of the lower fragment of the monument I; 𐰇𐰣𐰤 ~ üze ~ üze ~ upper, 𐰇𐰏𐰢𐰣𐰢𐰣 ~ tɯrime ~ teɯrim e ~ (“Oh, my Heaven!”), 𐰇𐰏𐰤𐰣 ~ YGz ~ yayiz ~ lower (дайр), 𐰇𐰏𐰣 ~ lma ~ elim e ~ (“Oh, my!”) My Country!, 𐰇𐰏𐰣 ~ sg ~ isig ~ one’s power, 𐰇𐰏𐰣𐰣 ~ brdma ~ berdim e ~ Oh I gave, on the 1st line of the west of the monument II; 𐰇𐰏𐰣𐰣 ~ WšGa ~ ušɣa ~ Ushga (2 times), 𐰇𐰏𐰤 ~ YQa ~ yaqa ~ edge (3 times), 𐰇 ~ B ~ ab ~ hunting, 𐰇𐰏𐰤𐰣 ~ YšQa ~ yišqa ~ at the mountain ranges, 𐰇𐰏𐰣𐰣𐰣𐰣 : 𐰇𐰏𐰣𐰣𐰣𐰣 : ~ :YWQ:tms: ~ yoq etmiş ~ sacrificed, mourned, 𐰇 ~ r ~ er ~ soldier, 𐰇𐰏𐰣𐰣𐰣 ~ trtma ~ ettirim ~ to organize, on the 2nd line of the upper fragment of the monument; 𐰇𐰏𐰤 ~ YQa ~ yaqa ~ edge, 𐰇𐰏𐰣𐰣𐰣 ~ WLGQa ~ uluɣqa ~ to Ulug on the 3rd line of the upper fragment of the monument. Through this frequency of merely three words, this monument has been inevitably viewed to be a quite special one in the source studies.

V. Conclusion

Most of the monuments were firstly found and studied by Ts.Battulga and R.Monkhtulga. That means no more researchers have studied and deciphered yet.

To say, as a conclusion for this presentation, based on the study, we should do any research particularly on lexicology and semantics of the runic inscription monuments found not only in Mongolia also in other countries, in a team but not a single or two person(s). Some alternative versions of deciphering any inscriptions can give a perfect result for further study. One example of it is to study and decipher the five characters incorporated in the inscription of Dalt. Those are unknown whether they are personal names or titles on which some researchers have proposed some versions of deciphering. As a result, there are no any way to include all the versions of its deciphering in this study, and we need to compile an ancient Turkic dictionary and to do some lexicological study in collaboration with any research team who specialised in linguistic, historical and cultural fields.

List of abbreviation

Asgt.	Inscription of Asgat
Bgt.	Inscription of Bugat
Deed.ts (Tsokh.)	Inscription of Deed Tsohiot
Del.u	Inscription of Del Uul
Don.sh	Monument of Dongoyyn Shiree
Dund.ts (Tsokh.)	Inscription of Dund Tsohiot
Dvrt.b	Inscription of Davirtiyn Buuts
Hirg.h	Inscription of Khirgis Khoology
Hirg.o	Monument of Hirgisiyn Ovoo
Htgr.h	Inscription of Khotgor Khag

Huls.a	Inscription of Khulsana am
Ikh.n	Inscription of Ikh Nart
Ivd.g	Inscription of Ivdey Deer stone
Khan.b	Inscription of Khanangiyn Buuts
Mng.a	Inscription of silver vessel
Tvt.u	Inscription of Tavit winter-camp
Uln.ch	Inscription of Ulaanchuluut
Urt.g	Inscription of Urtyn Gol

References

- Azzaya Badam. (2010). Moğolistan'daki Runik Yazıtlar. *Türkbilig, Türkoloji Araştırmaları*, 20, 67–81.
- Azzaya Badam. (2018). Moğolistan'da Bulunan Kayalar Üzerindeki Runik Yazıtlar(2010-2017). *Türkbilig, Türkoloji Araştırmaları*, 2018/35, 143–154.
- Battulga Tsend. (2016). Hirgisiin Hooloy Kazısında Bulunan Çatı Kiremidi Üzerindeki Yazıt/
Inscription on the Roof Tile in Khirgisiin Khoology Excavation. In *Moğolistan'daki Türk Ayak İzleri / Turkic Footprints in Mongolia* (pp. 124–130). Ulaanbaatar.
- Battulga Tsend. (2018). *A Newly-found Runic Inscription from Urtin Gol (Уртын голоос шинээр илрүүлсэн руни бичээс)*. 89–114. 천안: 단국대학교.

- Mönhtulga R. (2016). Moğolistan’da bulunan Eski Türk Dönemine ait Soğd tarzı gümüş kulplu tas / Sogdian-styled silver cup from Ancient Turkic period found in Mongolia. In *Moğolistan’daki Türk Ayak İzleri / Turkic Footprints in Mongolia*. (pp. 108–113). Ulaanbaatar.
- Mönhtulga R., & Ariyajav B. (2016). Moğolistan’da Bulunan Yeni Bir Yazıtı: Hulsana Am Yazıtı/ A Newly-Found Inscription from Mongolia: The Khulsana Am Inscription. In *Moğolistan’daki Türk Ayak İzleri / Turkic Footprints in Mongolia*. (pp. 85–87). Ulaanbaatar.
- Munkhtulga R. (2013). Silver Vessels from Ancient Turkic Period Found in Mongolia. *ACCU Nara International Correspondent. The Twelfth Regular Report, Vol. 12*, 27–32.
- Munkhtulga R. (2017). Ivdein Gol Deer Stone Runic Inscription. *Оюуны Хэлхээ, Боть XVI, Fasc.6*, 109–112.
- Sir Gerard Clauson. (1972). *An Etymological Dictionary of Pre-Thirteenth-Century Turkish*. Oxford, At the Clarendon Press.
- Аззаяа Б. (2016). “Жиримийн худаг”-ийн шинээр илрүүлсэн нэгэн бичээс. *Ази Судлал(ICAS 2016), Vol.II*, 576–579.
- Аззаяа Б. (2017). Монгол нутгаас шинээр илрүүлсэн руни бичгийн бага дурсгалууд. *ALTAICA, Vol.XIII*, 149–164.
- Аръяажав Б., & Аззаяа Б. (2014). Үхэр жил бичсэн эртний нэгэн бичээс (Урьдчилсан судалгаа). *Оюуны хэлхээ, I (11), (fasc.30,)*, 233–234.
- Баттулга Ц. (2010). Дэл уулын Жиримийн худагийн бичээсийн талаар дахин өгүүлэх нь. *Оюуны хэлхээ, Антоон Мостаэрт, Монгол судлалын төв, Боть II (07)*, 140–141.

- Баттулга Ц. (2013). *Өмнөговь аймгийн Номгон суманд ажилласан хээрийн судалгааны ажлын тайлан*. Улаанбаатар.
- Баттулга Ц. (2016a). Тайширын Давиртын бууцны бичээс. *Acta Historica, Tom XVII, Fasc.1*, 5–10.
- Баттулга Ц. (2016b). Хотгор хагийн Бийрэгийн бичээс. *Mongolian Journal of Anthropology, Archaeology and Ethnology, Official Journal of the National University of Mongolia, Volume 9 № 1 (471), December 2016*, 56–59.
- Баттулга Ц. (2017a, April 14). *Монголын Түрэг Судлал(2016). III*, 41–47. Улаанбаатар.
- Баттулга Ц. (2017b). Хиргисийн овооны гэрэлт хөшөөний бичээс. *ALTAICA, Vol.XIII*, 62–78.
- Баттулга Ц., & Аззаяа Б. (2009). Монголын руни бичгийн дурсгалууд. *Acta Historica, Tom.X(fasc.11)*, 78–96.
- Баттулга Ц., Анхбаяр Б., Cao Jian en, Батболд Г., Song Guodong, Амгалантөгс Ц., ... Li Chong lei. (2019). Цагаан туруут голын сав нутаг дахь эртний бичгийн дурсгалууд (Урьдчилсан судалгаа). *The 5th International Conference on Asian Studies, Vol.V*. Улаанбаатар: ШУТИС, Ази судлалын тэнхим
- Баттулга Ц., Бямбасүрэн Х., & Энхтөр Ч. (2010a). Тариатын өвөлжөөний түрэг бичээс. *Acta Historica, Tom.XI(fasc.15)*, 114–116.
- Баттулга Ц., Бямбасүрэн Х., & Энхтөр Ч. (2010b). Тариатын өвөлжөөний түрэг бичээс. *Acta Historica, Tom.XI(Fasc.15)*, 114–116.
- Баттулга Ц., & Идэрхангай Т. (2016). *Баянхонгор аймгийн нутагт хийсэн археологийн хайгуул судалгааны ажлын тайлан*. Улаанбаатар: МУИС, ШУС, Ази судлалын тэнхим, Түрэг судлалын салбар.

- Баттулга Ц., & Очир А. (2011). Хиргисийн хоолойн дөрвөлжингийн малтлагаас олдсон дээврийн ваар дээрх бичээс. *Acta Mongolica, Dedicated to the 100th Birthday of Professor F.W.Cleaves, Vol.11(366)*, 43–46.
- Баттулга Ц., & Цамбагарав Ж. (2015). Тавьтын өвөлжөөний эртний бичээс. *Mongolian Journal of Anthropology, Archaeology and Ethnology, Vol.8 No.1(403) December 2015*, 43–44.
- Болдбаатар Ю., Цэрэндагва Я., & Төрбат Р. (2017). Улаанчулуутын руни бичээсүүд. *Монголын Археологи-2016*, 74–80.
- Бүрэнтөгс Г. (2017). Умард Монголоос шинээр олдсон руни бичээстэй буган хөшөө. *Оюуны Хэлхээ, Боть XVI, Fasc.7*, 113–120.
- Идэрхангай Т., Баттулга Ц., & Баяр Б. (2017). Монгол нутгаас шинээр илрүүлсэн руни бичгийн дурсгалууд(Урьдчилсан судалгаа). *Studia Archaeologica, Tom.XXXVI, Fasc.15*, 231–238.
- Мөнхтулга Р. (2015, August 21). *Хулсана амны руни бичээс*. 45–46. Улаанбаатар.
- Мөнхтулга Р. (2018). *Түрэг, монгол судлалын өгүүлүүд, Эрдэм шинжилгээний бүтээлийн түүвэр (2003-2018 он)*. Улаанбаатар.
- Мөнхтулга Р., & Оосава Т. (2015). *Донгойн ширээн дурсгалын гэрэлт хөшөөний бичээсийг анхны удаа уншсан нь*. 21–55. Улаанбаатар.
- Наделяев В.М., Насилов Д.М., Тенишев Э.Р., & Щербак А.М. (Eds.). (1969). *Древнетюркский словарь*. Ленинград.
- Цэрэндагва Я., Шнейдер Ж., Розен А., Коннорс Р., Далантай С., Фаркухар Ж., ... Олзбаяр Г. (2016). *Монгол-Америкийн хамтарсан экспедицийн Их Нартын байгалийн нөөц газарт хийсэн археологийн хээрийн шинжилгээний ажлын тайлан (хээрийн*

судалгааны тайлан) [Хээрийн судалгааны тайлан]. Улаанбаатар: ШУА, Түүх-Археологийн Хүрээлэн.

Эрдэнэ М., & Баттулга Ц. (2012). Далтын хадны зураг, эртний бичээс(Урьдчилсан судалгаа). *Mongolian Journal of Anthropology, Archaeology and Ethnology, Vol.7, No.1(378) December 2012, 62–70.*

바트돌가 체. (2017). *새로 발견된 문자 문헌자료: 히르기스 오보 비석 연구 (Руни бичгийн шинэ сурвалж: Хиргисийн овооны гэрэлт хөшөөний бичээсийн судалгаа)*. 91–121.
천안: 단국대학교.

HISTORY AND DEVELOPMENT OF DICTIONARIES ON INDIGENOUS ENDANGERED LANGUAGES OF CENTRAL INDIA: THEIR PAST, PRESENT AND FUTURE

Mendem Bapuji

University of Hyderabad

Abstract

The linguistic diversity is a unique feature, strength, identity and back bone of India which deserves to be preserved by the preparation of dictionaries. A unique characteristic feature of Central India which comprises of Odisha, Chhattisgarh, MadyaPradesh and some parts of Maharashtra and Andhra Pradesh is unity with cultural and linguistic diversity. Among the above mentioned states, southern Odisha especially Koraput district is rich in accommodating languages of three different genetic families, viz. Dravidian, Indo-Aryan (Indo-European) and Munda (Astro-Asiatic). Among these languages, few of them are major and many of them are indigenous or tribal endangered languages. According to the report of the Council of Analytical Tribal Studies (COATS) situated in Koraput district of Odisha lists down the 62 indigenous tribal communities. These groups are divided into two groups namely primitive tribal groups (PTGs) latter designated as particularly vulnerable tribal groups (PVTGs) and non-primitive tribal groups (NPTGs). The history of preparation of dictionaries on these languages will be highlighted in the paper. Along with the history, the paper also tries to discuss at length the problems involved in the preparation of the dictionaries. Since most of the languages in the area are endangered at different levels (based on UNESCO's classification) what steps will be helpful in preparing the dictionaries for these minor languages and how ultimately these dictionaries would be helpful in preserving these endangered languages will be highlighted. The paper also tries to tabulate the number of dictionaries prepared for each language family and see the ratio of language endangerment family wise. Finally the paper attempts for the steps to preserve these languages from the lexicographic point of view.

Key Words: Central India, Linguistic Diversity, Endangered Languages, Indigenous and Dictionaries

Introduction

India is one of the countries which consist of more number of languages among which indigenous languages more in number. Most of these languages have flourished with rich and oral tradition since from the advent of the Aryans into Indian subcontinent. Lexicographical point of view, less number works were compiled comparing to the major and scheduled languages of India comparing to the indigenous languages. We do not have various types of dictionaries that meet different requirements for the indigenous languages. The existing dictionaries do not reflect the modern developments in the field of linguistics and lexicography.

Since India has become digital country, there is lot of requirement for the use of Indian languages in the field of education, administration, and mass communication etc. This situation demands for the need of compiling new dictionaries especially on indigenous languages.

The word Tribe and the condition of Indigenous Languages

As defined by Annamalai (2000) “Tribe” is an administrative term and legal term which used in the colonial period to label some ethnic groups based on their religious practices, customs and socio economic status in order to give special attention to them as mandated by the constitution of India”. Since the word arises many ramifications and perceptions, the word is not used in the paper. The word indigenous is used in the paper because it reflects the nativity and the origin of the communities. Indigenous languages are the languages which are native to a particular region with their own origin myths and indigenous knowledge. These languages in India are underdeveloped for several ages. Though they are rich in oral literature they were never considered as scheduled languages. In most of the cases they were and are being dominated by the major languages. These are the non-vehicles of power and prestige and never used as a medium of instruction, though the constitution of India provides article 350a which says all the children of the country should get his/her primary education in the mother tongue. These languages lack the scripts and in most of the cases they are confined mostly as home languages (which are used for the intra communication purpose). The literacy rate among the communities of these languages, never cross the 0-5% of the literacy. Even the census of India does not recognize them because most of the communities have less than ten thousand population. One good thing among the communities of these languages is illiterate bilingualism/multilingualism. Bilingualism among the communities of these languages is very common and in some places they are trilingual too Ramaiah & Reddy (2005, vol. VI, p-424). But these indigenous languages are endangered due many factors hence the use of language is curtailed and not passing to the next generations. Most of these languages are either on the verge of extinction or endangerment.

Linguistic Profile of India

India is a country of many languages. The linguistic profile of India varies about the number of languages spoken and the number of languages estimated. The tentative assumption made by the linguists in India about the language is 200 or so. Among these 200 languages 80% of the languages are indigenous, minor, tribal and neglected languages. Some of these 80% of the indigenous languages do not have even small glossaries of the languages. They were highly neglected and undermined by the governments and by the policy makers.

Tentative Family-wise number of languages (Reddy, 2003)

Family Affiliation	Indigenous /Non literary Languages	Major/Literary Languages	Total
--------------------	------------------------------------	--------------------------	-------

Dravidian	24	06	30
Austro-Asiatic	20	-	20
Indo-Aryan	20	30	50
Tibeto-Burman	100	-	100
Andamanese	10	-	10
Total	174	36	210

Linguistic Profile of Central India

Parts of Maharashtra, Odisha, Chattisgarh, and Madhya Pradesh are considered to be Central India. Central India is the place which hosts languages belonging to many family of languages viz. Dravidian, Indo-Aryan and Munda family of languages. Most of the languages found in the region are indigenous and non-literary languages. More Dravidian languages, More Munda languages are spoken in the area than the main belts (region) of the family languages. The following are some of the indigenous languages spoken in the Central India.

Dravidian Family of Languages	Availability of Grammar	Availability of Dictionary	Language Status
Gondi	Yes	NO	Vulnerable
Kudukh	Yes	NO	Endangered
Konda	Yes	NO	Endangered
Kui	Monograph	NO	Endangered
Kuvi	Monograph	NO	Endangered
Manda	Yes	Yes	Endangered
Parji	Monograph	NO	Endangered
Pengo	Monograph	NO	Endangered
Ollari Gadab	Monograph	NO	Endangered

Munda Family of Languages	Availability of Grammar	Availability of Dictionary	Language Status
Gutob Gadaba	Yes	NO	Endangered
Gorum	No	Yes	Endangered
Gta?	No	NO	Endangered
Juang	No	NO	Endangered
Jureyi	No	NO	Endangered
Karia	Yes	NO	Endangered
Remo	No	Yes	Endangered
Sora	Yes	YES	
Indo-Aryan Family of Languages	Availability of Grammar	Availability of Dictionary	Language Status
Bhill	Few works	NO Information	Endangered
Halbi	Few works	NO Information	Endangered

If we look at the above table, it illustrates that most of the languages in the Central India do not have dictionaries. Efforts were less in the making of the dictionaries on these indigenous languages. A documented language will never be extinct. It can be used even in the absence of the speakers, if the language is documented in the form a dictionary or a grammar. The Central Indian situation clearly depicts that there is a clear need in the preparation of the dictionaries before they shift the major languages. Due to the globalization, unemployment and various other socio cultural problems are leading these indigenous communities to shift from their mother tongues to major languages of areas. In the case of some languages of the Central India, the shift is taking place at the cost of native tongues. The native tongues are not passed on to the younger generations; hence the younger generation is being shifting to the major languages of the area. In order to stop this phenomenon, grammars and dictionaries should be produced which can be used for the primary education purpose where students enjoy the teaching in their own language.

History of Dictionaries on Indigenous languages of Central India

If we look at the Central Indian situation or the history of dictionary production on indigenous languages of Central India is very poor. Some of the missionaries from the 18th and 19th century have prepared some glossaries for these indigenous languages for the sake of understanding indigenous tongues. The first work that covered the indigenous languages is LSI which is often referred as Linguistic Survey of India started in 1894. The survey helped in identifying the many indigenous languages and language families of Central India and India. The survey has covered most of the languages with grammatical information and some folk stories from the communities with glosses. The description can be seen from the below picture.

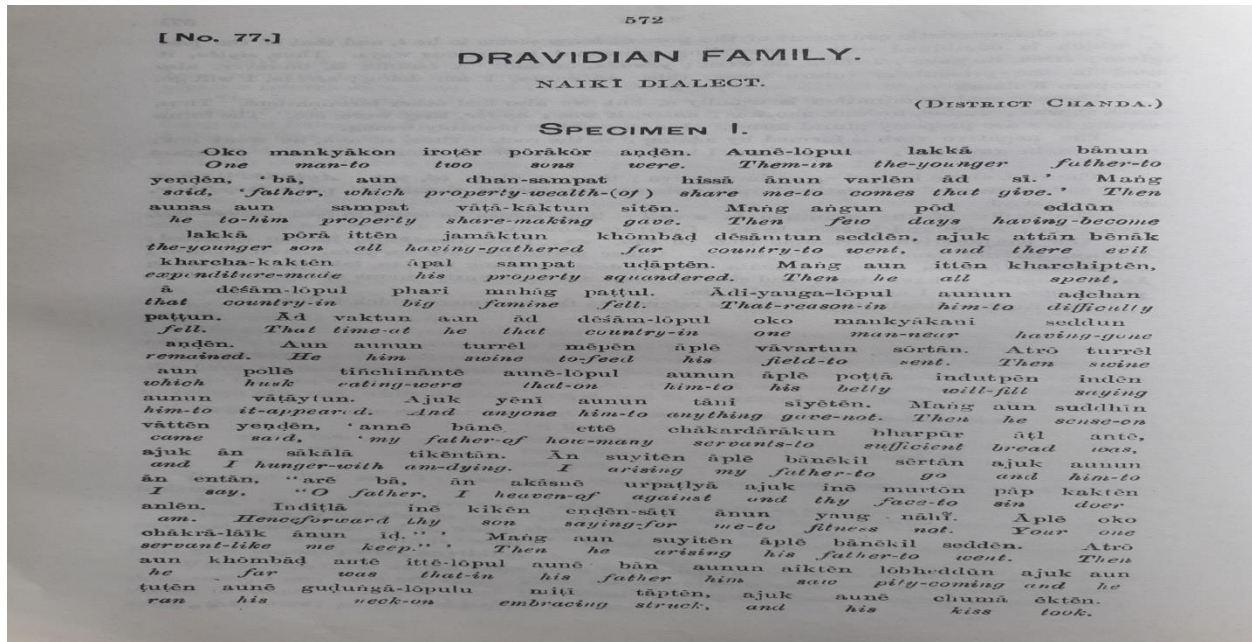


Figure 1 LSI Description of Folk Story from Naiki Language

Before the survey (LSI) there was no single dictionary on these minor languages. As they had enumerated many indigenous languages Anthropologists from various countries started working on these languages. Especially Anthropological Survey of India and Language division from the census of India have played a very vital role in studying the languages and the cultures of these indigenous languages. As they were studying the cultures of these languages they encountered that unless one can study the language it was impossible to understand the culture of the people, hence started describing the languages. It was the first step from the description point of view. As a part of the description, linguists and anthropologists used to give a list of words from the studied language ranging from 150-500 words. These words also random and there is no precise method was used in listing down the words. The following picture will illustrate the phenomenon:

Word list follows:

Sl.no.	English	Gadaba(Ollari)	Sl.no.	English	Gadaba(Ollari)
1	air	ua:l varida:	45	wife	asma:l
2	ashes	ni:D	46	woman	asma:l
3	cloud	mogul	47	ant	suidil
4	cold	pañil	48	bird	sitepa:p
5	darkness	sikka:	49	cat	verig
6	earth	tuku:D	50	claw	va:ndel
7	eclipse	goron	51	cock	ga:M'ja
8	fire	kis	52	cow	koNDe
9	fog	dumriandar	53	dog	M'ette
10	forest	koppel	54	egg	ga:r
11	hill	kupli	55	feather	ke:ndisil
12	ice	adir ka:nD	56	fish	niNil
13	moon	neliM	57	fly	mosi
14	mountain	ber koppel	58	fox	bokDa:
15	rain	ku:Di	59	goat	mege
16	river	pereD	60	grease	koDuku:T
17	road	berba:	61	guts	puDu:k
18	sand	maND	62	horn	korkusul
19	wind	besiva:l (storm)	63	back	po:Tiel
20	wood	kaDasil	64	belly	puDu:k
21	baby	sepal	65	blood	ba:ni
22	boy	sepa:l	66	body	men
23	bride	koDa:l/ile	67	bone	puNukul
24	bride groom	salgiND/ileND	68	breast	pa:lgin
25	brother (eld)	berba:i'	69	chest	argil
	(yong)	hamcepal	70	ear	kekol

Figure 2 Depicts the random list of words at the end of monograph from Ollari Gadaba language

Later several linguist from abroad and India and language lovers started working on these minor and neglected languages. The first work on one of the indigenous languages (Sora) was started officially by Ramamurthy in 1938. He prepared English Sora dictionary. Later people started preparing the dictionaries with the local scripts and the languages are from the indigenous communities. The following picture will illustrate the phenomenon:

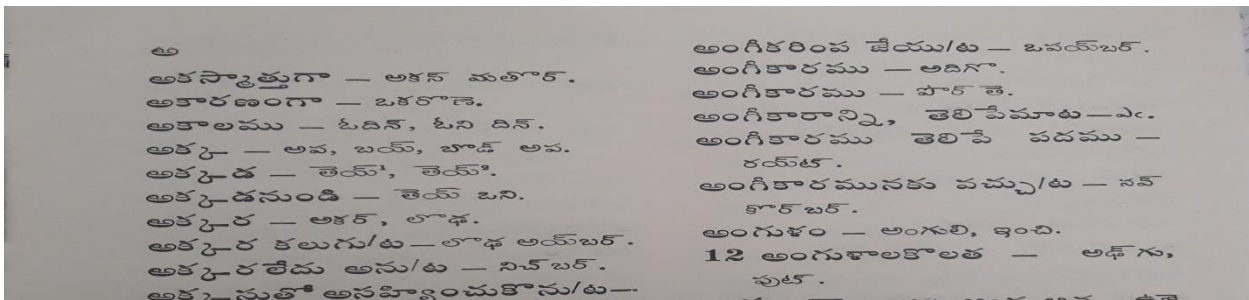


Figure 3 Telugu Adivasi dictionary with local scripts

Another step in the preparation of dictionaries on indigenous languages was preparation of dictionaries with pictures which are known as pictorial dictionaries. These dictionaries are prepared alphabetically in the local scripts along with the roman script and the pictorial illustrations. The following picture will depict the phenomenon:

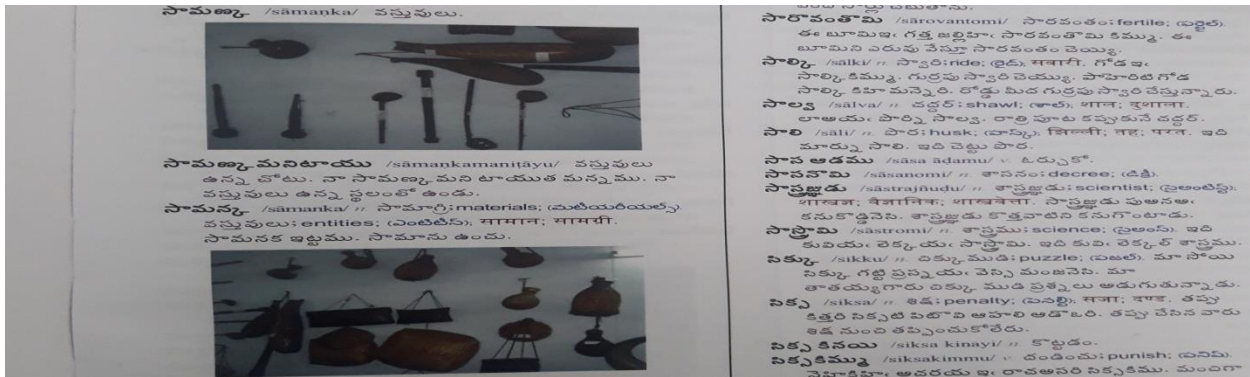


Figure 4 Depicts the making of dictionary with local scripts along with the Roman script and pictorial description

The second half of the nineteenth century is fruitful in the area of indigenous languages in which descriptive linguistics has its peak. As the enumeration and family affiliation was clear linguists started describing the languages with phonemic inventory, graphitization and orthography. As these things were advanced, word lists from these minor languages have got precise form with clear orthography and phonemes and with comparative vocabulary from the neighboring languages of the same family, at the end of the description of the each minor language. The trend has continued till 1980s and 1990s. The following picture will illustrate the phenomenon:

PART THREE

COMPARATIVE VOCABULARY

The alphabetical order adopted in the Vocabulary is as follows :

Vowels :—*a, ā, ā̄, ə, i, ī, u, ū, e, ē, o, ō, ǝ*

Consonants :—*k, g, ṅ, c, ts, j, dz, z, ṅ̄, ṭ, ḍ, ṇ, t, d, n, p, b, m, y, r, ṛ, l, v, s*

(Some Tamil words have been transcribed as they are pronounced.)

AGLE, *sb.*, pot

AṬ-, *vb.*, to strike ; to rain in torrents ; cf. De. *pāni mār-siyāse* 'it is raining heavily' [Ta. *aṭi* 'to beat' 'to strike', Ka. *oḍi* 'id.', Pj. *aṭṭ-* 'to strike', māva *aṭṭ-* 'to harrow', *poṭ-kul aṭṭ-* 'to clap' 'to snap fingers']

AṬAṬI-, *sb.*, fighting [duplication of *aṭ-* ('to strike') to denote reciprocity ; cf. Halbi *tapa-tapi* 'fighting', Beng. *mara-mari* 'id.', mā- 'to beat']

AND-, *vb.*, hunger or thirst is

ABA, pl. -*r*, *sb.*, father [Konḍa *eba* 'father' ; cf. Tibetan, etc. *apha*, etc.; Santali, Munḍari, etc. *aba* 'father'—used in the vocative by children ; Ta. Ma. *appan* 'father', Ka. *appa*, Tu. *amme*, Te. *appa*, *abba*, Kur. Brah. *abbā* 'id.']

AM-ABA, pl. -*r*, *sb.*, my-father (lit. our-father)

AM-AYA, pl. -*v*, *sb.*, my-mother (lit. our-mother)

AMB, pl. -*ul*, *sb.*, arrow [Ta. Ma. *ampu* 'arrow', Ka. *ambu*, Te. *ambu*, *ammu*, Pj. *amb*, Konḍa *am*, Kui *āmbu* 'id.']

Figure 5 Depicts the precise form of the word list with cognates from the neighboring languages of the same family

Later with the establishment of full pledged lexicography departments in the respective universities of the Indian Union, dictionary making has become part and branch of applied linguistics by extending the grammatical information, person, number and gender information's. The use of phonetic and phonemic scripts also got prominence in the preparation of the dictionaries. In this direction mono/bi/and multilingual dictionaries were produced for a few indigenous languages of Central India. The major achievement in period is preparation of the etymological dictionary for all the Dravidian language by Thomas Burrow and M.B Emeneau in 1894. This was a robust work for the major as well as minor languages of the different family of languages. The work helped in establishing the family relations, proto form the family and in comparative linguistics. The following picture will illustrate the phenomenon:

142 *Ta.* attan father, elder, person of rank or eminence; attai, attaicar father's sister, mother-in-law, woman of rank or eminence; attān elder sister's husband; father's sister's son, maternal uncle's son when elder, wife's brother when elder; attācci elder brother's wife, husband's sister; attimpēr elder sister's husband; father's sister's husband; atti elder sister; attō excl. of wonder; tattai elder sister. *Ma.* atta mother, mother's sister; attan father. *Ka.* atte, atti mother-in-law; sōdar-atte, sōdar-atti father's sister, mother's brother's wife; attike elder sister; attige elder brother's wife. *Tu.* attè mother-in-law, aunt; attigè elder brother's wife. *Te.* atta mother-in-law, father's sister, maternal uncle's wife. *Nk.* atiak (pl. -ev) father's sister. *Nk. (Ch.)* ato bāy id. *Ga.* (Oll.) āta, (S.³) atta id. *Go.* ātī father's sister (*Voc.* 127). *Kui* ata, atali grandmother. *Kuwi* (S) atta aunt; (Isr.) atu grandmother. / Cf. Skt. attā- mother, mother's sister, elder sister (*lex.*); atti(kā)-, anti(kā)-, artikā- elder sister (*lex.*); Pkt. attā- mother, mother-in-law; father's sister's husband; Turner, *CDIAL*, nos. 221, 222. *DED(S)* 121.

Figure 6 A Word from DED with Cognates from other Dravidian Languages

In the preparation of the etymological dictionary, 28 Dravidian languages were taken into consideration. Tamil which is considered as a post proto language after the protolanguage started separating into different languages is taken as a basic primary language and the rest of the languages. Alphabetical order is followed in the dictionary. Roman script used for the orthography purpose. The dictionary does not contain the proto-Dravidian (PDr) reconstructions. Most of the materials used in the dictionaries from the works of various linguists in the field, who have collected data from the extensive field observations. The total number words used in the dictionary are approximately 5000 words with cognates.

Sarva Siksha Abiyan Contribution in the preparation of Dictionaries

To fulfill the constitutional right of education provided by Indian constitution, article 350a, **Sarva Siksha Abiyan** which is present called as **Rajiv Vidya Mission** started preparing dictionaries especially for the indigenous languages. The main aim of the project is to promote mother tongue use, mother tongue education at the level of primary education and finally to preserve the indigenous languages from the state of language endangerment. The dictionaries in the project are prepared based on the semantic domains of the languages with the local scripts of the states rather than the meaning wise. This project provided education right for the indigenous children to undergo his/her primary education in their own mother tongue rather than state official languages. In this project some of the Central Indian languages like **Kuvi**, Gondi and some other Dravidian languages have got the chance to get the dictionaries. These dictionaries even

helped in the preparation of primers for the **Kuvi** and Gondi children. The following picture will depict the phenomenon:

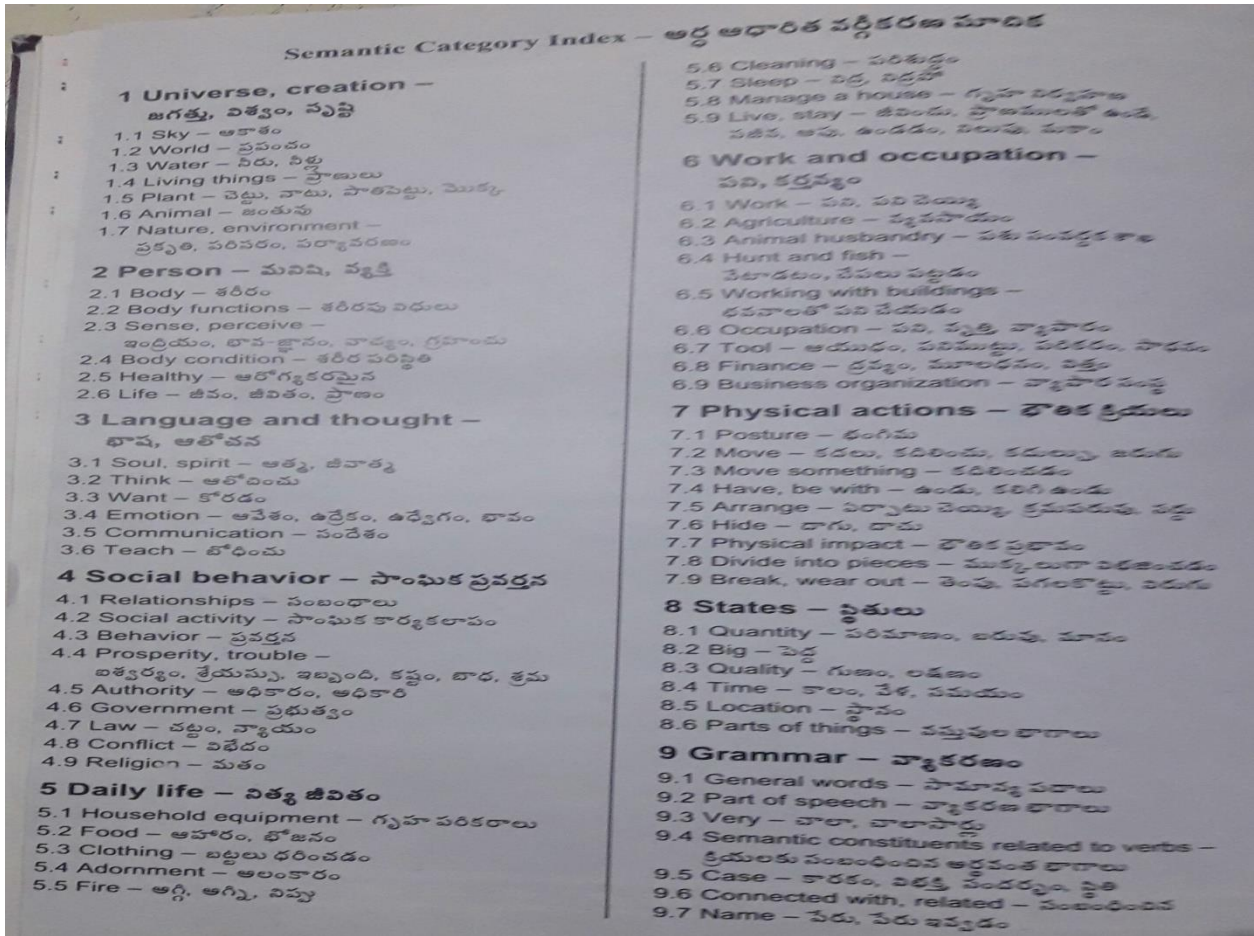


Figure 7 Kuvi dictionary with semantic domains

Apart from **Rajiv Vidya Mission, CIIL** (Central Institute of Indian Languages) also played a vital role in the preparation of dictionaries for the regional languages. Since from the inception in 1969, the institute has been working on Indian regional languages persistently. The institute has trained a number of linguists in the field of lexicography for the preparation of dictionaries especially for the indigenous languages. Its contribution was enormous for Indian regional languages.

The second Paradigm shift in the preparation of dictionaries in India is Computational approach. In this approach digital dictionaries are being prepared with computer aid and with different tools developed various institutions like SIL (Summer Institute of Linguistics) etc.

Draw backs advantages in the preparation of dictionaries for indigenous languages

If we look at the history of dictionaries on the indigenous languages of the Central India, most of them have vocabulary lists, glossaries as an appendix to grammars. Never full pledged dictionaries were prepared on

these languages. Reddy (2003) states that, dictionaries tries to encapsulates the socio cultural way of life in its words and phrases – reflecting the categorization of physical and spiritual world as perceived by the community. The dictionaries prepared n these languages will reveal the language system to the outside world and preserves the language for posterity. When dictionaries are prepared on these languages the constitutional right og imparting primary education will be fulfilled. Along with this the dictionaries also will be helpful in teaching and learning the indigenous languages as a medium of instruction or as a second language. This will pave the way to translation tribal literature and the language to other Indian languages and vice versa. Reddy (2003:4) discusses a cases study of **Kondh** community in which a Festival called **Tuki** is celebrated every year. In the festival lexical items that reveal about religious functionaries, official’s viz. priests, sacred specialists, dignitaries involved in the sacrificial festival will be used. For example: Reddy (2003:4)

“mur	‘high priest, the chief’	ja:ni	‘prest’
gurma:y	‘the woman possessed with spirits’	dihavari	‘astrolger’
du:tar	‘representatives on the festival committee’		
mudradarya	‘the person who takes the sacrificial animal around the village’		
ra:pya	‘ the man who scavenges the blood and flesh of sacrificial animal’”		

There is a need for giving more importance in the dictionary preparation is that, the culture loaded lexical items are structure in their connotative and denotative meanings. These words give the clear picture of the culture of the ethnic group. Another important thing one has to take into consideration that, inclusion of binomials in the preparation of dictionaries on these indigenous languages.

Binomials

Reddy (2003) states that a binomial (BN) or a lexical doublet (LD) is a set o two words in a certain order whose members are of an identical syntactic category, pertain to a selected semantic field which exhibit a specific sense-relation between them, and (they) may sometimes be connected by a lexical link. These are not same as onomatopoeia, reduplication and echo-formation and other expressive both in form and function. There is constraint on the use of these binomials. They cannot be reversible. If at all, it happens to be reversed they demand a kind of conjunction of two items but not as an (BN). Each language has its own way of forming the binomials and some languages allow reversibility. Sometimes these binomials demand an identical category restriction on the two members like noun-noun, adjective-adjective and verb-verb constructions. Some the examples are taken from Ollari Gadaba a Dravidian language can be seen below. Some examples are taken from the Bapuji (2019) description of Ollari Gadaba language.

wal	wa:in	‘air and rain’
aba	si:nɔ	‘father and son’
narka:m	pakta:l	‘day and night’
aya	aba	‘mother and father’
ta:li	gina	‘rice and plate’

Indigenous languages of Central India in particular and general in India are very rich in expressivities especially binomials. The use of these binomials indicates an idiomatic native expression system and the cultural life of the communities. Hence these also should be given utmost care in the preparation of the dictionaries.

Conclusion

Government India has started many projects for the preparation of dictionaries and translation through various projects under Ministry of Communication and Information Technology, Technology development for Indian languages. The projects are The Indian Languages Indian Languages Machine Translation Project (ILIL), English Language to Indian Languages Project (ELIL), Indo-Word NET project and Cross Lingual Information Access Project. Under these projects dictionary preparation also one of the aims in the projects. None of the projects have included any of the indigenous languages of India. This is a major drawback for indigenous languages. If at least a few of the languages were included, it would have been a great resource for these languages. Only Central Institute of Indian Languages (CIIL) is doing under the scheme called Scheme for Protection and Preservation of Endangered languages (SPPEL). Finally I conclude that indigenous languages which rich in oral literature and spontaneous poetry should be given much importance through the preparation of Dictionaries and grammars so that these languages can be preserved from the language endangerment. Preservation is better step than the revitalization.

References

- Annamalai, E. 2000. “The Linguistic Heritage of India” in Koul, O.N. & L. Devaki (Eds.), *Linguistic Heritage of India and Asia*. Mysore: Central Institute of Indian Languages. 1-6.
- Bhattacharya, S. 1956. *Ollari, a Dravidian Speech*. Memoir No. 3. Delhi: Manager of Publications.
- Bhattacharya, S. S. 2002. “Languages in India: Their Status and Function” in S.H. Itagi & S.K. Singh (Eds.), *Linguistic Landscaping in India*. Mysore: CIIL Publication, 54-97
- Burrow, T and M. B. Emeneau. 1984. *A Dravidian Etymological Dictionary*. Oxford: Clarendon Press.
- Burrow, T. and Bhattacharya, S. 1960. *Comparative Vocabulary of Gondi Dialects*: Journal of the Asiatic Society.

- Census of India 2001. *Paper 1 of 2007: Language*. New Delhi: Office of the Registrar General, India.
- Emeneau, M. B. 1969. The Non-Literary Dravidian Languages. In Sebeok, T. S. et.al. (ed.). *Current trends in Linguistics*. The Hague: Mouton.
- Grierson, G.A. 1904. Linguistic Survey of India: Muṇḍa and Dravidian Languages. Vol-4. Delhi: Low Price Publications.
- Halliday, M.A.K. and Yallop Colin. 2007. *Lexicology: A Short Introduction*. Wiltshire: Cromwell Press.
- Katre, S. M. 2003. Current Trends in Indian Lexicography. *Lexicography. Critical Concepts II*, Londres-Nueva York, Routledge, 147-157.
- Krishnamurti, Bh. 1969. Koṇḍa or Kubi; A Dravidian Language. Hyderabad: Tribal Cultural Research & Training Institute.
- Krishnamurti, Bh. 2003. *Dravidian Languages*. Cambridge: Cambridge University Press. .
- Linguistic Survey of India Special Studies, Orissa. 2002. Kolkota: Language Division, Office of the Registrar General, India.
- Mahapatra, K. 1976. Echo Formation in Gtaq. In: Jener, T.ans S. Starosta (eds.). *Austro AsiaticStudies*. Honolulu: University of Hawai Press.
- Misra, B. G. (Ed.). 1980. *Lexicography in India: Proceedings of the First National Conference on Dictionary Making in Indian Languages, Mysore, (Vol. 4)*. Central Institute of Indian Languages: Mysore.
- Ramakrishna, B. Reddy. B. 2003. *Bilingual Dictionaries for Unwritten Languages*. An unpublished Research Paper. Osmania University.
- Ramaiah L.S. and Ramakrishna Reddy, B. 2005. *Tribal and Minor Dravidian languages and LinguisticsAn International Bibliography of Dravidian languages and Linguistics series Volume VI*. Chennai: T.R. Publications.
- Ramakrishna, B. Reddy. B. 2013. "The Tribal Languages of Odisha". *International Journal of Dravidian Linguistics*. 42. 40-62.
- Ramammurti, G. V. 1938. *English Sora Dictionary*. Madras: Government Press.
- Rameshkumar, K. 2018. *Remo/Bondo Pictorial Dictionary*: Mysore: Central Institute of Indian Languages.
- Reddy, J. 1979. *Kuvi Grammar*. Mysore: Central Institute of Indian Languages.
- Singh, R. A. 1982. *An Introduction to Lexicography*. Mysore: Central Institute of Indian Languages (CIIL).

- Subrahmanyam, P. S. 1968. *A Descriptive Grammar of Gondi*. Annamalai Nagar: Annamalai University.
- Suresh, J. 2001. *Gadaba (Ollari)*. Linguistic Survey of India Orissa: Calcutta.
- Thurston, E and Rangachari, K. 1909. *Castes and Tribes of Southern India*, Vol. 1-7, Government Press, Madras.
- Thusu, K. N. and Jha, M. 1972. "Ollar Gadaba of Koraput", *Anthropological Survey of India* 27, Calcutta.
- Winfield, W. W. 1929. *A Vocabulary of the Kui Language Kui English*, Calcutta, Asiatic Society of Bengal.
- Zide, N. H. and dass, B.P., 1966. *Gutob Verb Lexicon*. Mimeo: Chicago.

Online References:

- http://www.censusindi.gov.in.census_Data_2001/Census_Data_Online/language /data_on_language.html
- http://www.ethnologue.com/ethno_docs/introduction.asp
- <http://www.unesco.org/new/en/culture/themes/endangeredlanguages/atlas-of-languages-in-danger/>

THE METHOD OF THE REAL LIFE BASED SCHOOL DICTIONARY*

Bülent Özkan

Mersin University

Ferdi Bozkurt

Anadolu University

Nurettin Demir

Eskişehir Osmangazi University

Erdoğan Boz

Hacettepe University

Şükrü Halûk Akalin

Hacettepe University

Abstract

The aim of The Real Life Based School Dictionary Project (RLBSD) is to introduce a new, original and real life based school dictionary as an educational material. The current project can be qualified as a research project encompassing formal and informal learning for all educational environments in developing competencies indicated in the Turkey Competencies Framework. In line with the method implemented, the expected outcome of this project is to create a real life based school dictionary as an instructional material supported with information technology. The lack of a corpus-based school dictionary for Turkish in the literature proves that the existing dictionaries include dictionary entries based on old methods and intuitions of native speakers of the language and that they were developed with a non-experimental approach. The project team and the staff have an interdisciplinary qualification in line with the aims and objectives of the project. The project team will include researchers and staff from different disciplines such as linguists, lexicology experts, educational scientists, Turkish language experts, software experts, field teachers and graphic designers. In addition, the proposed project foresees the participation of all stakeholders in the RLBSD creation process such as field instructors, students and linguistics / lexicography specialists. In the current study, lexicographical the method of real life based school dictionary will be presented.

* This Project is funded by, TÜBİTAK SOBAG - 1003. Project No:118K109. We thank TÜBİTAK for the contributions.

Key Words: Lexicography, method, school dictionary, corpus linguistics

1. Introduction

Corpus linguistics, as is known, is a discipline that can present -especially in lexicography studies- real-life and experimental outputs via texts chosen from natural language environment. This feature enables corpus linguistics to provide the most suitable and the most contemporary opportunities for the researchers in creating a real life based school dictionary.

The lack of a corpus based school dictionary and existing school dictionaries' including entries based upon the intuitions of first language speakers mean that these dictionaries were not created with experimental approaches. Within this scope, *Real Life Based School Dictionary (RLBSD)* meets a significant educational material need.

It is unanimously accepted view that it is necessary to take the needs of the target group to the forefront while creating a dictionary. It is also necessary to conduct a needs analysis in the lights of scientific and experimental methods (Atkins et al., 1995; 85). It is acknowledged that the applications and methods that have been used in creating Turkish general and specific dictionaries such as picking up witnesses, relying on personal knowledge, creating token indexes etc. are of no function in creating dictionaries.

The intuitions and personal choices of linguists and lexicographers lead lexicographers to different subjective results far from scientific facts (McEnery, 2006: 145). Modern lexicography creates dictionaries with an approach based on experimental findings and language use, and thanks to linguistic databases - corpora- that can represent the language itself. To sum up, today dictionaries are created, parallel to information technologies, on platforms that can produce extremely efficient outcomes both for the users and the lexicographers.

1.1. Objective

The purpose of the study is to create a new, original and *real life based* school dictionary as an instructional material. Therefore, the main objective of the study is to build a special purpose corpus by collecting the natural language use examples that the students can encounter during their education life from existing course books, approved by the Turkish Ministry of National Education, children's literature works, periodical children's publishing, etc., and to create the *Real Life Based School Dictionary (RLBSD)* based on the corpus.

The base research question of the study can be posed as *What does the vocabulary of linguistic environment that school-age children encounter include?*

1.2. Scope of the Study

The scope of the study is to reveal a new, original and real life based school dictionary as an instructional material based on a special purpose corpus which includes the natural language use examples such as existing course books, children's literature works, periodical children's publishing, etc. that the students can encounter during their education life (from primary school to high school) by following the principles and methods of corpus linguistics.

The sub research questions within the framework of the main research question *What does the vocabulary of linguistic environment that school-age children encounter include?* are as follows:

1. *What does the vocabulary of course books approved by the Turkish Ministry of National Education include?*
2. *What does the vocabulary of natural linguistic environments that the children encounter include?*
3. *What are the effect(s) and efficiency of the RLBSD on children's vocabulary?*

2. Methodology

The methodology in determining the vocabulary of the RLBSD can be summarised as **1.** Needs analysis via focus group discussions, **2.** Entry selection from the special purpose corpus that will be built for the RLBSD, **3.** Creating the RLBSD within the scope of linguistics and lexicography, **4.** The evaluation of the RLBSD's efficiency in educational environment.

These stages are explained in detailed below:

1. Focus Group Discussion and Needs Analysis

In qualitative studies, observation, focus group discussion and document analysis can be used to collect data. In this study **Focus Group Discussion (FGD)** is one of the data collection tool in determining the needs to create the RLBSD. The purpose of FGD is to collect qualitative data from the perspectives, experiences, interests, tendencies, thoughts, perceptions, attitudes and habits of the participants on the predetermined topic (Stewart and Shamdasani, 1990; Kitzinger, 1994, 1995; Krueger, 1994; Gibbs, 1997; Bowling, 2002).

FGD is frequently used in developing instructional materials (McBrien, Felizardo, Orr and Raymond, 2008) and in educational research (Byers and Wilcox, 1988; Barbour and Kitzinger, 2001; Gizir, 2007). FGD is to obtain information in detail, to generate ideas, to use the effect of group dynamics, and an interview between a small group and a leader (Bowling, 2002). FGD is defined as a carefully designed discussion environment where the participants can freely utter their opinions (Krueger, 1994). In this sense, the first stage of creating the RLBSD is to determine the needs. FGD was conducted with the teachers, stakeholders

of the dictionary, and the lexicographers and the details of lexicographic structures were determined. [Focus Group Discussion is summarised from Çokluk, 2011.]

2. The Principles and Methods of Lexicography and Corpus Linguistics in Creating the RLBSD

This stage of the study was designed in descriptive and relational model. While descriptive research aims to determine the situation as it is, relational research is used in studies in which there is no cause-effect relationship, in studies in which no or partial manipulation and control can be done by nature or because of practicality (Erkuş, 2009; Karasar, 2009; Büyüköztürk et al., 2010). Specially designed data collection tools, techniques or methods are used to collect data (Erkuş, 2009). In this study, *the corpus linguistics methods and techniques that enable to reach experimental results* were used.

The corpus used to create the **RLBSD** is a **specialized corpus** as defined in the literature. The specialized corpus comprises of two main layers. **The first layer of the corpus** comprises of a specialized corpus, Turkish Child Literature Corpus (TCLC), created to analyse Turkish Children's Literature qualitatively and quantitatively. TCLC was built within the Tübitak-Sobag-1001 project. In that project, the texts were selected from Turkish Children's Literature in terms of its *primary vocabulary, the readability and age suitability, lexical diversity and lexical semantic patterns, morphological, lexical and syntactic features, and internal and external structural features of the children's literature texts*. **TCLC**, including **8.639.522** lexical items and 1089 different text types, will be used as lexicographic database in the **RLBSD**.

The second layer of the corpus comprises of the course books approved by the Turkish Ministry of National Education. The course books were digitalized and added to the corpus according to the corpus building phases and principles.

2.1. The Morphological, Lexical and Syntactic Features of the Corpus created for the RLBSD

The corpus was created by parsing the texts; classifying each sentence under the main layers or sublayers [*Turkish Children's Literature* (novel, poem, story, periodicals etc.), *MNE approved course books* (in terms of grade and course), *Other Types* (other texts that children may encounter)]; by tagging the metadata such as the author's name, the title of the text, publication year etc.

Lemmatization and deduplication processes followed the abovementioned phase of corpus building process. The entries in the dictionary are based on this deduplication process.

Each lexical item in the corpus was analysed with morphological analysis tool developed by Kemal Oflazer (*bk.* <http://www.hlst.sabanciuniv.edu/TL/>).

2.1.1. Sentence and Sample Sentence Tagging

< *Mısır piramitlerinin yüksekliklerini hesaplayan ve "bir dairenin içine dik üçgen çizen" büyük bir bilim adamıydı Tales.* >

<Çocuk Yazını Derlemi (Children Literature Corpus)> **1. layer**

<Deneme (Essay)> **text type**

<mavisel yener> **the author's name**

<Pramitler (Pyramids)> **the title of the text**

<2009> **publication year**

2.1.2. Sample Morphological/Attributional Analysis

Mısır [???] | *pramitlerinin* [piramit+Noun+A3pl+P3sg+Gen] | *yüksekliklerini* [yüksek+Adj^DB+Noun+Ness+A3pl+P3sg+Acc] | *hesaplayan* [hesapla+Verb+Pos^DB+Adj+PresPart] | *ve* [ve+Conj] | *bir* [bir+Det] | *dairenin* [daire+Noun+A3sg+P2sg+Gen] | *içine* [iç+Noun+A3sg+P3sg+Dat] | *dik* [dik+Adj] | *üçgen* [üçgen+Noun] | *çizen* [çiz+Verb+Pos^DB+Adj+PresPart] | *büyük* [büyük+Adj] | *bir* [bir+Det] | *bilim* [bilim+Noun] | *adamıydı* [adam+Noun+A3sg+P3sg+Nom^DB+Verb+Zero+Past+A3sg] | *Tales* [???]

2.1.3. Data Processing Stages of RLBSD

An online platform is used for data processing. (<http://derlem.mersin.edu.tr/okulsozluk/>) (*Figure 1*).

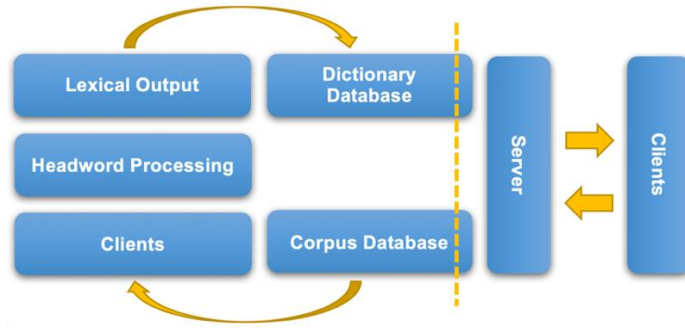


Figure 1: *Data processing platform* (Taken from Özkan, 2013).

Two databases of data processing platform work continuously to enable the users to use or to check data. To put it simply, the users will save the lexical queries they made on CORPUS DATABASE to DICTIONARY DATABASE. The processed data are shown instantly with query screens via SERVER.

2.1.4. The Data Processing Stages of the RLBSD on Corpus

The RLBSD is subjected to search, tagging and reporting processes over a system as seen in Figure 2.

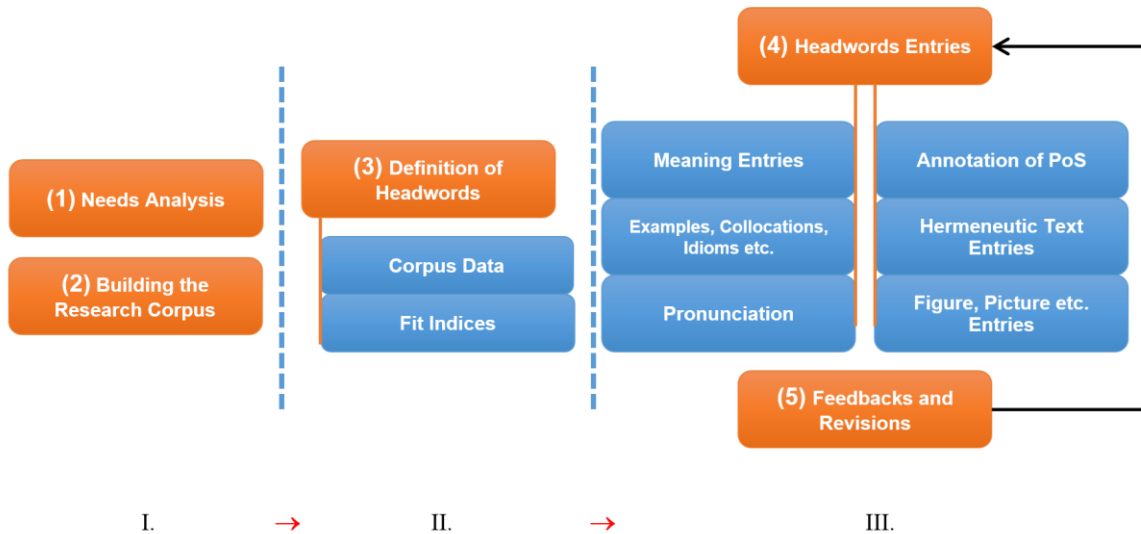


Figure 2: *Data processing stages*

It can be said that the work comprises of three stages and five main operations.

I. (1) needs analysis and (2) building the research corpus,

II. (3) determination of headwords (with the corpus data and fit indices)

III. (4) configuration of the headwords (entry entering, part of speech tagging, witness selection, determining the collocations and idiomatic expressions, pronunciation, reader friendly text and picture selection etc.) and **5) feedbacks and revisions.**

The explanations of data processing stages parallel to methods and techniques that will be used in the project are as follows:

(1) Needs Analysis:

In Lexicography the concepts of *school dictionary*, *children's dictionary*, *college dictionaries* (Hartmann and James, 1998: 122) are used in similar studies and these correspond to “a dictionary written for school age children”. The common features of these dictionaries are including controlled set of vocabulary and having a simplistic design with visuals. Herein, the vocabulary of a school dictionary, word list, should be systematically determined. The entries forming the word list of a dictionary are called lexemes.

User research should reveal what the expectations of the users are and what is practical in a dictionary for the users. Jackson (2002: 163) emphasises that before the planning stage, it is necessary to define the target user group. The target user group can be determined via various ways such as field studies, surveys, observations, or expert opinions (Atkins and Rundell, 2008: 30). The target user group, their age range, educational level, and their purposes affect everything related to the dictionary.

In sum, it is one of the most significant factors to determine the needs of the target user group in writing a dictionary. Therefore, in order to determine the needs of the users in the proposed project, **Focus Group Discussion** was conducted with field teachers and linguistics/lexicographers.

(2) Creating the Corpus:

The corpus used in the RLBSD is a *specialized corpus*. There are two layers of the corpus. **The first layer of the corpus** is TCLC comprising of 8.639.522 (+/-) lexical items. **The second layer of the corpus** comprises of the course books approved by the Turkish Ministry of National Education. The course books were digitalized and added to the corpus according to the corpus building phases and principles.

(3) Determination of the Headwords:

When the history of lexicography is reviewed, it is seen that the linguistic materials used by the lexicographers in their studies are mainly based on L1 intuitions (armchair lexicography) and therefore these studies are full of mistakes and far from validity as they cannot reflect the linguistic reality precisely. Moreover, the intuitions and personal choices of linguists and lexicographers lead lexicographers to different subjective results far from scientific facts (McEnery, 2006). Today it is acknowledged that the

applications such as picking up witnesses, relying on personal knowledge, creating token indexes etc. are of no function in creating dictionaries (Atkins et al. 1995). On the other hand, school dictionaries are regarded as significant references in teaching first language and there are studies on how to write school dictionaries in the literature (Malkiel, 1967; Bergenholtz, H., and R. H. Gouws. 2012).

There are two different ways in determining the headwords in a dictionary. **The first way** is to obtain headword lists from the lexical items of a corpus. This stage is one of the standard data processing stages.

The vocabulary items of MNE approved course books that can be used as headwords in the **RLBSD** were determined according to the *fit indices* calculations. The opinions of three field teachers for each course book in addition to the frequency and distribution scores of the items were taken into consideration.

Besides, as the headwords would also be determined from the corpus, the vocabulary items obtained from the *fit indices* calculations and corpus were configured as headwords.

Hence the sources of headwords are: **a.** headwords from the corpus, **b.** headwords from the *fit indices* calculations.

(4) Configuration of the Headwords:

There are some specific headword configuration items in writing a dictionary. These are: *spelling, pronunciation, inflections, part of speech, meaning(s), definition, samples, usage, run-ons, and etymology* (Jackson 2002; Hanks 2003). Some further configurations can also be added to these thanks to information technologies (picture, audio file, etc.). Which one of these will be used is shaped by the aims of the lexicographer. The data obtained from the corpus form the basis of this shape.

The biggest advantage of corpora in linguistics and lexicography studies is the contributions they make to usage and experimental based decision mechanisms in determining the elements such as *context sensitive* meaning, part of speech, multiword items (collocation and idiomatic expressions). Therefore, according to the query results, the linguist/lexicographer can be flexible and generative in determining the meaning, part of speech and multiword items.

On the other hand, although school dictionaries structurally resemble to general dictionaries (headword, meaning, part of speech, etc.), they are to include configurational changes suitable to target group. The sources used in the **RLBSD** are listed below (Table 1).

a. Meaning/Definition configuration

b. Pronunciation

c. Determining PoS and conceptual field

d. Witness choice

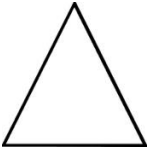
e. Determining the collocations and idiomatic expressions

f. Explanatory text configuration

g. Configurations of figures, pictures, etc.

h. Vocabulary teaching, Word games, Spelling tests

Table 1: Sample Headword Configuration [*Üçgen (triangle) headword sample*]

Headword	üçgen
Meanings	Üç açısı, üç kenarı olan geometri biçimi.
Pronunciation	üçgen
Part of Speech/Usage	ad. önad. geometri
Witness	<ul style="list-style-type: none">• <i>Mısır piramitlerinin yüksekliklerini hesaplayan ve "bir dairenin içine dik üçgen çizen" büyük bir bilim adamıydı Tales. [Çocuk Yazını Derlemi]</i>• <i>Martın boynunda asılı duran üçgen biçimindeki metal parçayı da o zaman fark etti. "Boynundaki nedir?" diye sordu. [Çocuk Yazını Derlemi]</i>•
Collocations and Idioms	üçgen piramit üçgen prizma çeşitkenar üçgen dik üçgen eşkenar üçgen
Hermeneutic Text	Figure, Picture etc
Açılara Göre Üçgenler Açılarına göre üçgenler üç çeşittir. <i>Dar Açılı Üçgen:</i> Bütün açıları 90 dereceden küçük (dar açı) olan üçgene dar açılı üçgen denir. <i>Geniş Açılı Üçgen:</i> Açılardan biri 90 dereceden büyük (geniş açı) olan üçgene geniş açılı üçgen denir. Bir üçgende yalnız bir açının ölçüsü geniş açı olur. <i>Dik Açılı Üçgen:</i> Açılarında biri 90 derece (dik açı) olan üçgene dik açılı üçgen denir. Bir üçgende yalnız bir açının ölçüsü 90 derece olabilir. [Matematik 5.Sınıf Ders Kitabı]	
Vocabulary Teaching, Word Games, Spelling Tests	

(5) Feedbacks and Revisions:

The project is interactive and has many stakeholders. Only the opinions of lexicographers are not enough to create a school dictionary. The usage and likes of the users are also included to the project as one of the stages in creating as feedbacks and revisions. In this sense, the necessary revisions and corrections can be done thanks to the feedbacks of students, lexicographers and field teachers.

3. The Evaluation of the RLBSD's Efficiency

The other data collection tool of the study is a survey. Surveys are a kind of *semi-structured written interview technique and/or tool* used to collect data. As being relatively easy to prepare and the possibility to reach many participants in a short time, surveys are widely used in social sciences. Surveys are used to describe an existing phenomenon (Erkuş, 2013:161, Büyüköztürk, 2016:124). In the study, the evaluations of the RLBSD will be collected through the survey from lexicographers, teachers and students.

On the other hand, in order to find out the efficiency of the RLBSD in different grades, two groups (one experimental, one control) from the 5th grade and two groups (one experimental, one control) from the 8th grade were compared (each grade was compared in itself) in terms of practicality of a printed dictionary and the RLBSD. Firstly, a **reading comprehension test** will be implemented to form homogenous groups based on the reading comprehension level. The students in the control group will be given a printed dictionary, and the students in experimental group will use the RLBSD. The students will be compared in the processes of text analysis and text comprehension whether there will be a significant difference between the groups using the RLBSD and printed dictionary.

Conclusion

As a result, the stages to be followed in creating the RLBSD **1.** determination of vocabulary/needs through focus group discussion, **2.** selecting the lexical entries from the special field corpus to be built for the RLBSD, **3.** creation of RLBSD within the framework of corpus linguistics – lexicography, and **4.** evaluation of the effectiveness of the RLBSD in educational environment correspond to a practice not previously experienced in Turkish Lexicography.

Since the distribution of the RLBSD to be created through the methods to be implemented will be carried out via information technologies, the outputs of the study can be used throughout the country. In this sense, the RLBSD will take its place in educational environment as an effective course material with its quality as it will be accessed easily through web, tablet, smartphone and smart board applications.

Since the RLBSD offers an opportunity for a flexible and updateable information platform, it also bears the qualification of being a sustainable model. In addition, how RLBSD affects learning outcomes can also be tested because its effectiveness will be evaluated.

References

- Atkins B. T. S. & Beth L. 1995. "Building on a Corpus: A Linguistic and Lexicographical Look at some Near-Synonyms". *International Journal of Lexicography*. Vol. 8 no. 2: 85-114. Oxford University Press.
- Atkins, B. S., & Rundell, M. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.
- Barbour, R.S. & Kitzinger, J. (Eds.). 2001. *Developing Focus Group Research: Politics, Theory, and Practice*. London: SAGE.

- Bergenholtz, H., and R.H. Gouws. 2012. What is lexicography? *Lexikos* 22: 31–42.
- Bowling, A. 2002. *Research Methods in Health: Investigating Health and Health Services*. Philadelphia, PA: McGraw-Hill House.
- Büyükköztürk Ş. vd., 2010. *Bilimsel Araştırma Yöntemleri*. Ankara: PegemA Yayıncılık.
- Byers, P. Y. & Wilcox, J. R. 1988. “Focus groups: an alternative method of gathering qualitative data in communication research”, Annual Meeting of the Speech Communication Association, 74th, New Orleans, LA, November 3–6, 1988.
- Çokluk, Ö. 2011. Nitel Bir Görüşme Yöntemi: Odak Grup Görüşmesi. *Kuramsal Eğitim Bilim*, 4 (1), 95-107, 2011.
- Erkuş, A. 2009. *Davranış Bilimleri için Bilimsel Araştırma Süreci* (İkinci Basım). Ankara: Şeçkin Yayıncılık.
- Gizir, S. 2007. “Focus groups in educational studies”. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 3 (1), 1–20.
- Hanks, P. 2003. *Lexicography. Computational Linguistics*. (Ed. Ruslan Mitrov), Oxford University Press.
- Hartmann, R.R.K. 1998. *Dictionary of Lexicography*. London-New York: Routledge.
- Jackson, H. 2002. *Lexicography: An Introduction*. Routledge. USA. 169-171.
- Karasar, N. 2009. *Bilimsel Araştırma Yöntemi*. Ankara: Nobel Yayın-Dağıtım.
- Kitzinger, J. 1994. “The methodology of focus groups: the importance of interaction between research participants”, *Sociology of Health and Illness*, 16 (1), 103–121.
- Kitzinger, J. 1995. “Qualitative research: introducing focus groups”, *British Medical Journal*, 311, 299–302.
- Krueger, R.A. 1994. *Focus Groups: A Practical Guide For Applied Research*. London: SAGE. Gibbs, A. (1997). “Focus groups”, *Social Research Update*, 19.
- Malkiel, 1967. *School Dictionaries for First Language Learners*. *Lexikos* 22: 333-351.
- McBrien, S., Felizardo, G.R., Orr, D.G. & Raymond, M.J. 2008. “Using focus groups to revise an educational booklet for people living with methicillin-resistant”, *Health Promotion Practice*, 9 (1), 19–28.
- McEnery, T. vd. 2006. *Corpus-Based Language Studies -An Advanced Resource Book-*. London: Routledge.
- Özkan, B. 2013. Yöntem ve Uygulama Açısından 'Türkiye Türkçesi Söz Varlığının Derlem Tabanlı Sözlüğü'. *bilig. Türk Dünyası Sosyal Bilimler Dergisi* 66(3): 149-178.

Stewart, D.W. & Shamdasani, P.N. 1990. Focus Groups: Theory and Practice. Newbury Park, CA: SAGE.

<http://derlem.mersin.edu.tr/okulsozluk/>

<http://www.hlst.sabanciuniv.edu/TL/>

BERBER LEXICOGRAPHY: SEMANTIC AND MORPHOLOGICAL PROBLEMS

Carla Ferrerós Pagès

Universitat de Girona

Abstract

This study presents theoretical and methodological issues related to the development of Berber dictionaries, which have been largely discussed among scholars studying that language (Bonfour et al., 1995). Apart from certain problems such as the lack of monolingual dictionaries, imbalance in the description of different geographical varieties, and the lack of pan-Berber lexicographical sources, we will focus on a fundamental matter of interest. Traditionally, the Berber language has been analyzed morphologically like other Afro-Asiatic languages, isolating consonantal roots. This kind of analysis has affected lexical arrangement in dictionaries and has created problems for authors and users as, due to diachronic evolution, many Berber words cannot be analyzed and lexicographical works may be insufficient for non-linguistic users. The aim of this study is, first, to offer a panoramic view of existing lexicographic resources in that language and, second, to present an analysis of the problems related to root arrangement in Berber dictionaries in order to take them into account in future research. This analysis will be done from a morphological point of view, but our discussion will also consider how the type of arrangement in a dictionary can be related to semantic concerns such as linearity in the classification of meanings. Our results will show that lexical arrangement by consonantal roots involves certain problems in Berber lexicography and we will also suggest (in a preliminary way) that an adequate semantic structuration based on cognitive theories on polysemy can facilitate some matters related to this kind of arrangement. Berber lexicography has largely depended on European lexicography and Arabic contact. As such, this paper aims to take a step further in Berber lexicographical research.

Key Words: dictionaries, lexicography, semantics, morphology, Berber, use

Introduction

As Bonfour et al. (1995) show, “la lexicographie berbère est tributaire des langues européennes et une recherche en langue berbère semble à l’ordre du jour. [...] Mais l’etatut socio-politique du berbère pèsera encore lourdement sur la recherche.” Some of the problems mentioned by Bounfour et al. (1995) in their article on Berber lexicography refer to the lack of monolingual dictionaries, imbalance in the description of different geographical varieties, and the lack of pan-

Berber lexicographical sources. However, our paper will focus on describing a problem that we consider fundamental, apart from those mentioned above: the consonantal roots order in many Berber dictionaries.

It should be noted that the Berber language is an Afro-Asiatic language and has traditionally been analyzed, from the morphological point of view, like other languages of the same family, such as Arabic (Boukhris et al, 2008; Lamuela, 2003; Kossmann, 2000; Múrcia, 2011; etc). That is, considering that word roots are consonantal and that vowels and other consonants have morphological information. Inflected and derivative word forms are the result of processes of prefixing and suffixing other vocalic and consonant elements, but also of vowel organization modification and the root consonant quality (Ferrerós, 2014). However, there are also unpredictable phenomena that can make detecting roots difficult. These difficulties are usually caused by phenomena explained by a diachronic language study. Thus, for example, the diachronic loss of radical consonants leads to the existence of biconsonantal or uniconsonantal roots, which is why there is a large number of homophonic roots in many Berber dialects unlike other languages with triconsonantal roots such as Arabic. On the other hand, the unpredictability of nominal derivative morphology complicates root detection (leaving apart nominal or verbal inflective morphology and verbal derivative morphology, as they are more predictable) (Chaker, 2013).

The hypothesis from which we start in this preliminary study is twofold. First, we consider that lexical arrangement by consonantal root is not entirely adequate in Berber lexicography, especially concerning nominal items. In any case, although this type of arrangement presents some problems, the Indo-European type of arrangement would not be problem-free either, taking into account Berber morphological particularities. The other hypothesis from which we start, and considering that tradition has imposed the former kind of arrangement, is that a careful semantic analysis (within a cognitive semantics framework) would contribute to overcoming the semantic challenge posed by root arrangement. In Indo-European languages such as English, in which each entry usually corresponds to a word that belongs to a single grammatical category, the same lexical item usually has many related meanings. Those meanings are multiplied in Afro-Asiatic languages, as many different grammatical categories correspond to the same root and therefore a careful semantic analysis is necessary.

The purpose of this study is also twofold: in the first place, we will offer a current view of Berber lexicography. In the second place, we will describe and question, by means of the morphological analysis of Berber lexical items, the lexical arrangement by consonantal roots. Finally, we will show, in a preliminary way, how semantic study can be useful in lexicography in relation to the problems previously described.

This is an initial research on several topics related to Berber dictionaries. Due to the multiplicity of topics covered, it is a general survey of some problems related to Berber lexicography rather than an in-depth study. We will deal with the described issues in a preliminary way with the purpose of using them as a starting point for future research.

Method

This is a descriptive study. We will focus on analyzing, from different points of view (listed below), some recent Berber dictionaries (1980-present). We have made the selection taking into account Berber dialectal fragmentation:

- Dallet, J.M. (1982). *Dictionnaire kabyle-français, parler des AT MANGELLAT (Algérie)*. SELAF, Paris.
- Taifi, M. (1991). *Dictionnaire tamazight-français (parlers du Maroc central)*. L'Harmattan-Awal, Paris.
- Naït-Zerrad, K. (1998-2002) *Dictionnaire des racines berbères (formes attestées)*. Éditions Peeters, Paris-Louvain. 3 vol., up to <g>.
- Serhoual, M. (2002) *Dictionnaire tarifit-français*. Thèse de doctorat d'Etat dès lettres. Université Abdelmalek Essaadi, Tétouan.
- Múrcia, C. and Zenia, S. (2015). *Diccioniari català-amazic / amazic-català (estàndard del diasistema amazic septentrional)*. Llibres de l'índex, Barcelona.

From each dictionary, we will briefly analyze the following matters related to the objectives described in the Introduction:

- We will determine whether the dictionary is monolingual or bilingual and, in the latter case, which languages it includes. The content will be analyzed taking into account some of the shortcomings in earlier Berber lexicography, described by Bonfour et al. (1995). Then, the systematic presence of certain matters will be evaluated: if the dictionary contains a list of abbreviations and a list with the spelling of phonetic transcriptions, if it identifies and characterizes the dialect that it describes, if it mentions the sources used, etc.
- From each dictionary, the lexical arrangement by consonantal roots will be analyzed. Because of the scope of this task and because it is an introductory study, this matter will be illustrated through two examples:
 - o A uniliteral root will be selected in order to compare the entries in all the dictionaries to analyze the homophony issue mentioned in the Introduction.
 - o A derived word will be chosen and it will be compared among dictionaries to assess whether certain morphological features can be analyzed and whether the dictionaries take it into account. That is, we will determine whether the consulted dictionaries include the derived word in the entry headed by the simple word root or include it in a separate entry.

First, we will briefly expose the results obtained from the lexicographical sources observation (Section 3). Next, in the discussion section, the obtained results will be analyzed in order to verify

or refute the hypotheses set out in the Introduction. Finally, we will present the conclusions we have reached through the results analysis (Section 5).

Results

1. Description of dictionaries

Before listing the general characteristics of the dictionaries, we will make some important remarks. The Berber language has a large dialectal fragmentation and there is often no mutual understanding among dialects. In consequence, it is pertinent to note that dictionaries usually describe a certain variety. Dallet (1982), Taïfi (1991), and Serhoual (2002) describe, respectively, the Kabyle variety or Taqbaylit (Algeria), the Tamazight variety (Central Morocco), and the Riffian variety or Tarifit (Northern Morocco). The other two dictionaries are different in this regard. Naït-Zerrad (1998-2002) compiles, in three volumes (at the moment up to <g>) all the roots seen in different Berber varieties with the intent of doing extensive descriptive work. Múrcia and Zenia (2015) aim to offer a standard variety of Berber taking into account the varieties spoken in Morocco and Algeria, where the greatest number of Berber speakers live.

Tots els dialectes amazics són continuadors d'una protollengua comuna, però la varietat de referència –estandarditzada composicionalment atenent els diversos blocs dialectals– que ha de permetre la vertebració de l'immens espai comunicatiu amazic encara està en procés de codificació i no coneix més que una implementació incipient. (Múrcia and Zenia, 2015: xviii)

Some important issues related to the standardization process of the Berber language are present in the dictionaries we analyze. We discuss them below:

- There is no agreement in the use of some spellings to transcribe the same sound. For example, the sound [x] is transcribed as <x> in Dallet (1982), in Naït-Zerrad (1998-2002), and in Múrcia and Zenia (2015). On the other hand, it is transcribed as <ḥ> in Taïfi (1991) and <ḥ̣> in Serhoual (2002).
- There is no agreement between the uses of a phonetic or phonological type of spelling. For example, the word for “eye” is transcribed in the following ways, with an alternation of *ɣ* (phonological choice) and *ʔ* (phonetic choice, by assimilation of *ɣ* to feminine morpheme *ʔ*):

<i>tɪʔ</i>	root ʦ	Dallet, 1982
<i>tɪʔt</i> < <i>tɪʔt</i>	root ɣ	Taïfi, 1991
<i>tɪʔt</i>	root ɣ	Naït-Zerrad, 2002
<i>tɪʔ</i>	root ʦ	Serhoual, 2002
ʔɣEʔ (tɪʔt)	root E (ɣ)	Múrcia and Zenia, 2015

- There is no agreement in the use of a certain writing system. Múrcia and Zenia (2015) use the Tifinagh alphabet. All other dictionaries use the Latin alphabet.

Leaving aside matters related to the current process of language coding and standardization, which is not the goal of this study, the following table shows the characteristics of each dictionary:

	Abbreviations, signs, and symbols list	Spelling and transcription list	Dictionary languages	Dialect description	Sources description	Other
Dallet (1982)	Yes	Yes	Kabyle-French (1 direction) ¹⁰	Yes: Kabyle	Yes: lexicographic sources and survey	Root organization justification. Entries organization description. Dictionary content description. Annexes: relationship between Kabyle dictionaries and authors, verb conjugation and pronoun table, Kabyle names list, and illustrated sheets.
Taïfi (1991)	Yes	Yes	Tamazight-French (1 direction)	Yes: Tamazight	Yes: lexicographic survey	Kabyle and Arabic comparison. Root organization justification. Entries organization description.
Naït-Zerrad (1998-2002)	Yes	Yes	Berber-French (1 direction)	Yes: different varieties	Yes: bibliographic list	Dialectal phonetic differences description. Entries organization description.

¹⁰ A second volume published in (1985) constitutes a French-Kabyle appendix.

Serhoual (2002)	Yes	Yes	Riffian-French (1 direction)	Yes: Riffian	Yes: lexicographic survey	Root organization justification. Entries organization description. Notation system explanation.
Múrcia and Zenía (2015)	Yes	Yes	Berber-Catalan / Catalan-Berber (bidirectional)	Yes: compositional standard	Yes: lexicographic and grammatical sources	Introduction: Berber language, lexical stratification, dictionary characteristics, user instructions. Grammatical summary. Onomastic annex.

Table 1: comparison of recent Berber dictionaries

2. Lexical arrangement

Next, we present the collected data that will allow us to analyze certain problems related to the organization by consonantal roots used by all the consulted dictionaries. On the other hand, as indicated in Section 2, we describe how the dictionaries compile entries by uniliteral roots in order to later relate it with the homophony caused by diachronic loss of root consonants.

Each dictionary compiles a large number of homophone entries in the same F root:

- Dallet (1982): 16 entries
- Taïfi (1991): 14 entries
- Naït-Zerrad (1998-2002): 41 entries
- Serhoual (2002): 12 entries
- Múrcia and Zenia (2015): 7 entries

The homophony issue is not an anecdotal fact. It is common when it comes to mono-consonantal (and biconsonantal) roots. Thus, for example, if we take another mono-consonantal root like M (which does not appear in Naït-Zerrad, 1998-2002) we find similar data to those we have just shown:

- Dallet (1982): 18 entries
- Taïfi (1991): 17 entries
- Serhoual (2002): 36 entries
- Múrcia and Zenia (2015): 11 entries

Another issue related to the root arrangement problem is the one that refers to the treatment of some derived words, especially those that have few productive morphemes (that is, setting aside inflective morphology, verbal derivative morphology, and some cases of nominal derivative morphology).

We use, first, a simple pan-Berber word found in all the consulted dictionaries: the word for “nose.” It is seen under the root NZR or NSR in all cases (except in Naït-Zerrad, 1998-2002, as it is an unfinished work):

- Dallet (1982):

NZR

◆ *enzer* ; v. *enser*, n s r, F. III, 1355, *enher* ; 1419, *insiren*, *sinser* *inezzer* ; ur *yenzir -anzar* || Se mou-cher. || Avoir un grand nez.

◆ *tinzer* (ti) ; F. III, 1354, *tenhert tinzar* (ti) / *tanzarin* (ta) || Narine. Nez. || Amour-propre. Honneur. • *bu tinzer* / *bu tanzarin*, individu au petit nez. • *laz ur yesei tinzar*, la faim n'a

NZR

_____ *gunzer* / *kunzer*, *Izy* / *munzer*, *Izd* *gunzer*, *tgunzur*, ur-*gunzer* = saigner du nez
S_____ *sgunzer* / *skunzer* / *smunzer* *sgunzer*, *sgunzur*, ur-*sgunzer* = faire saigner du nez (en donnant un coup), donner un coup sur le nez.

- Taïfi (1991):

- Serhoual (2002):

NZR

◆ *finzar*, nfp., nez. V. *nsar*.

NSR

◆ *nsar*, vt. ; *insar*, wa *ynsar*, *inessa*, ad *insar*, tz. ; qv *enser*, *nesser*. || Se moucher. V. *fenzar*.
N *kunzar*. ◊ ad *inessar yah̄tur di zzi' a t yssird* : il se mouchera dans un mouchoir et il le lavera. ◆ *ansar* faire saigner du nez recipr. ; se donner recipr. des coups sur le nez.

_____ *tinzer* (tn)

tinzar (tn) = narine. Pl.: nez • *qqent'-as tenzar*, il a les narines bouchées.

_____ *tigenzer* / *tiyenger* (tg)

tigenzar (tg) = mm. ss. q. précéd. = maille.

◆ *inzer* (yi) ; (*anzaren*, v. le suiv.) || Nez ; p. j. au sg., peu empl.

◆ *anzaren* (wa) / *tanzarin* (v. ci-dessus).

|| Nez. || Honneur. Amour-propre (*nif*). • *bu wanzaren*, individu au nez trop gros. • *tiferrawin bb^manzaren*, les ailes du nez. les narines. • *tumatac m man-*

Nœud. Boucle, AH

_____ *anzaîn* (wa), sans sg.

= nez (un ou plusieurs). • *tuffin-as wanzaîn*, il a un gros nez. • *nzeîn-as wanzaîn*, il a un nez droit et fin. • *bu-wanzaîn*, qui a un gros nez.

_____ *tanzarin*, sans sg.

= petit nez, nez d'enfant (un ou plusieurs).

_____ *agunzer* (u)

igunzerîn = saignement de nez.

_____ *asgunzer* (u)

isgunzurîn = action de faire saigner du nez, de donner des coups sur le nez.

_____ *asenzar* (u)

isenzarîn, *taenzart*, *tienzarin* (te) = individu au nez trop court ou qui n'a pas de nez. = nasilleur, qui parle du nez.

_____ *tienger* (te)

= fait d'avoir le nez trop court. Nasillement.

Múrcia and Zenia (2015):

lʒo NZR

- *nprim* † **lʒo** † (†l-) ~ *pl* † **lʒo** † *anat nariu* (*en sg*), *nas* (*en pl*) † **lʒo** † *té el nas tapat* || *fig* amor propi, honor ♦ TRG *tenzärt* ~ *pl tinzar*, YDMS *ānzəy tənzart* ~ *pl tənzar* idem
- *nprim* † **lʒo** † (lʒo-) ~ *pl* † **lʒo** † *nariu* (*en sg*), *nas* (*en pl*) || *fig* amor propi, honor | *dim* † **lʒo** † (†o-) ~ *pl* † **lʒo** † *nasset*

(w-), na ; qr *anser*. ♦ *anzar* (wa-), nms., pl. *anzaren*; zn. bq. Am. sj. gz. pl. *anzaran*; qr. *inzar*, pl. *anzaren*. || Nez, gros nez (symbole de l'amour-propre). ♦ gz. *anzaran ines d izegraren*: il a le nez long, son nez est long. ♦ Loc. *issek *ires deg wanzarn*. ♦ Loc. *anzarn-ines d isemmađen* (ou *asemmađ n wanzarn*): son nez est froid, il est pusillanime, veule. V. *mfes. aqarran/qarn*. ♦ *tinzar* (tn-), nfp., pl. *anzarn, tz.*; W. *tinzar*; bt. bq. *tinzart*; zn. *tinzerđ*, pl. *tinzarin*. || Nez, tz. W. bq. bt. V. *ayenzur*; orgueil (du noble), par ext. || Narine, gz. zn. bq. Am. sj. ♦ *tit n tinzart*: narine. ♦ *issawar s tinzar*: il parle du nez, il nasille. V. *nefuef*. ♦ Loc. *aqšur n tinzar*: morve sèche du nez; vaurien, ordure (pers., fig.). ♦ Loc. *if-ij zi tinzar ad iyennej*: prends-le par le bout du nez, il chantera, il dira tout ce qu'il a sur le cœur. ♦ Loc. *issek *ires deg wanzarn*. ♦ *bu-wanzaren*, ams., pl. *ayđ bu-wanzaren*, fém. *mu-wanzaren*, pl. *suyđ mu-wanzaren*. || Qui a un grand ou un long nez, celui ayant un nez camus; camard. ♦ *bu-wanzaren irebzen*: celui dont le nez est aplati.

We see the roots under which words derived from *anzar* are compiled as follows:

- Taïfi (1991) compiles the following information under the root GNZR:

___ gunzer < anzar^*n*, nez. Cf. n·z·r
 tgnzur = saigner du nez.
 ___ tigenzert / tiyenzert (tg)
 tigenzar = narine = maille, noeud. Cf. n·z·r

- Naït-Zerrad (2002) compiles the following information under the root GNZR:

CHL V00 wwunzer, gunzer “saigner du nez” // N1,N3 awunzer « hémorragie (par le nez) » V10 zwunzer
 MC V00 gunzer, yunzer, kunzer, munzer « saigner du nez » ; tigenzert, tiyenzert,

« narine ; maille, noeud » ; pl. *tiynzar, tigenzar* « nez » [...]

- Serhoual (2002) compiles the following information under the root γ NZR:

aynzur (u-), nms., pl. *iyznurn* [...]. Visage, figure (sale, laid, de mauvaise mine, péj.) [...] Visage, joue [...].
Nez V. *tinzar-nsar* [...] Pommette du visage [...] Laideron; malchanceux.

And under the root QNSR (not as clearly related, but possibly):

aqensur (u-), nms., pl. *iqensuren, iqensar* [...] Visage, figure, péj.

- Múrcia and Zenia (2015) compile the following information under the root XNZR (XlʒO):

vprim – [...] *estar refredat, estar constipat*

nest XlʒO med. *Refredat, constipate, cadarn* [...]

Discussion

1. Description of dictionaries

The linguistic description of Berber has its origins in the nineteenth century, during French colonization (and Spanish colonization in the Rif area). The first dictionaries available were, above all, by European authors. Bonfour et al. (1995) distribute the history of Berber lexicography into three stages:

1. Utilitarian lexicography (1820-1918, pre-colonial stage): these works are aimed at merchants, travelers, the army, etc. There is little theoretical and methodological basis. They are bilingual or trilingual and they do not take into account dialectal variation.
2. Dialectal lexicography (1918-1950, colonial stage): systematic studies of dialects, with more theoretical basis on morphological and phonological language structure. The articles are more structured and the data have been extracted by the use of ethnographic surveys and texts. Some theoretical and methodological problems have not yet been addressed.¹¹
3. Scientific lexicography (1950-present, post-colonial stage): systematic studies of dialects. The articles are structured and there is a major consistency in phonetic transcription. They include

¹¹ Dictionaries such as Faucauld's, which contain a very complete lexical description of Tuareg, are from this second stage and are still used today, not only for their quality, but also because of the lack of more recent works of this variety.

sociolinguistic and linguistic introductions and the entries contain vast grammatical information. The articles are introduced by consonantal roots, unlike some works from previous stages.

In this study, we have analyzed five dictionaries that belong to this last stage. As Bonfour et al. (1995) point out, there were some problems related to Berber lexicography as of the beginning. The authors summarize them in the following points:

- There is inequality in the description of dialects. The best described ones, lexicographically speaking, are Kabyle and Tuareg. Even so, the most important work of this last variety is the four-volume dictionary written by Charles de Faucauld, dating back to the mid-20th century.
- The fact that there is no equal description of all varieties means that there is a need to work out a pan-Berber lexicon research that will lead us to some diachronic and semantic problem issues.
- Berber lexicography is dependent on European linguistics and contact with the Arabic language.

Bonfour et al. (1995) listed these problems before the end of the twentieth century. However, they have not been addressed to date. It is true that, on the one hand, some varieties that were not well described in 1995 now have new dictionaries available, such as the Riffian variety with the very complete dictionary by Serhoual (2002). On the other hand, we currently have an important, yet unfinished pan-Berber lexical source (Naït-Zerrad, 1998-2002) and a source that compiles more than one dialect with an aim at standardization (Múrcia and Zenia, 2015).

The five dictionaries we have described also make up for some of the shortages from previous period resources: they offer, systematically, all the information lacking in previous dictionaries, such as a phonetic or phonological transcription list of the spelling used in the work, or abbreviations and signs lists. In addition, the articles contain coherently organized grammatical information. All the works described in this paper provide other types of information: descriptions of the dialectal variety, sources of information gathering, and methodological matters regarding the dictionary's development. In some cases, they also offer maps (Dallet, 1982; Múrcia and Zenia, 2015), graphs, or drawings for ethnographic explanations (Dallet, 1982; Taïfi, 1991) and, as seen in Múrcia and Zenia (2015), a very complete grammatical summary.

However, some unsolved problems are listed below:

- There is no agreement in spelling use. The Berber language has just begun the coding process so, although in some cases the written tradition means that some spellings have been consolidated, there is no agreement with regard to some others, as we have seen in the previous section with the transcription of [x]. Between the five dictionaries, it is transcribed three different ways: <x> is the most frequent, and it seems that it is the recently agreed upon way for the transcription of this sound. However, we also find <ĥ> and <h>.
- In some cases, transcription follows phonetic criteria and, in others, phonological criteria. In the previous section, we have seen the case of words for “eye”: the root consonant, *d*, is phonetically

assimilated to the consonant of feminine affix *t...t*. Some dictionaries show this assimilation and others use a phonological writing. This fact can condition, as in that case, the root detection. That is why some dictionaries place the word for “eye” under the root *Ḍ* and others under the root *Ṭ*.

- There is not agreement in the writing system use: most lexicographical sources use the Latin alphabet, except one. Currently, in Morocco, the official alphabet to transcribe the Berber language is Tifinagh, and that is the option used by Múrcia and Zenia (2015), in a dictionary that aims at standardization. In spite of this, there is tradition in the use of the Latin alphabet. Many Berberologists consider the choice of the Latin alphabet as a better option and they would be in favor of reserving Tifinagh for symbolic uses and identity purposes. The debate remains open (El Aissati, 2014).
- One of the major problems, also identified by Bonfour et al. (1995), is the fact that there are no monolingual dictionaries. This issue has not been addressed between 1995 to the present.
- In all dictionaries, authors have seen the need to justify a root-based arrangement, which in some cases implies the addition of a grammatical summary to introduce users to some morphology notions in order to be able to use the dictionary. We will cover this question in the next section.

2. Lexical arrangement by consonantal roots

One of the most important issues in the development of Berber dictionaries is lexical arrangement. This paper describes some of the problems that arise. We have already mentioned that in Afro-Asiatic languages, roots usually consist of three consonants, and vowels and other consonants offer morphological information. Let us see the examples below:

iarya	“(it) is burned”
isurya	“(he) burns” (transitive)
tiyaryart	“heat, oven”

These three Berber words share the same root, *rɣ*, with a semantic value related to fire. In all three cases there are affixed inflective and derivative morphemes, which are constant and have a predictable meaning. However, in Berber, unlike languages such as Arabic, in many cases words are not so easily analyzed. Below is some data from Chaker (2013), described in greater detail by Ferrerós (2014), from the Kabyle variety:

- a. A 35% of the vocabulary is morphologically analyzable.
- b. A 35% of the vocabulary is morphologically analyzable by dialectal and diachronic comparison.
- c. A 30% of the vocabulary are non-analyzable isolate forms.

According to Chaker's (2013) data, roots from a third of the words can be easily isolated. Another third is analyzable by linguists with diachronic and dialectal variation knowledge. The remaining third cannot be analyzed. Chaker (2013) also claims that, due to diachronic issues, many root consonants have been lost and that is why it is common to find biliteral and uniliteral roots, which causes homophony. The previous section showed some data we will reintroduce here:

- F root

Dallet (1982): 16 entries

Taïfi (1991): 14 entries

Naït-Zerrad (1998-2002): 41 entries

Serhoual (2002): 12 entries

Múrcia and Zenia (2015): 7 entries

Two issues arise from this data. First, there are numerical differences between dictionaries. It is true that Naït-Zerrad (1998-2002) compiles the most F roots and this could be explained because it is a work that aims to compile all forms from all dialects. However, when it comes to data referring to the M root, we see that the Tarifit variety dictionary (Serhoual, 2002) compiles many more roots than the others:

- Dallet (1982): 18 entries

- Taïfi (1991): 17 entries

- Serhoual (2002): 36 entries

- Murcia and Zenia (2015): 11 entries

This numerical imbalance surely indicates the authors' difficulty in deciding whether roots are homophones (and therefore compiled separately) or if they are polysemous (and therefore compiled in the same entry). This leads us to the second issue, namely, what happens with two words derived from the M root in the Serhoual dictionary (although this matter is not exclusive to this author). Serhoual compiles in two separate (homophone) entries, with five other entries in between, the word for *ymma* "mother" and *uma* "brother," although he points out that *uma* is a compound composed of *u-* "son" and *-ma* "mother." There is the possibility that a dictionary user may have some difficulty finding those words due to doubts he or she may have regarding whether they are included or not in the same entry. On the other hand, the Serhoual dictionary dedicates ten-and-a-half pages to the M root. As a result, users, when searching for a word, also encounter difficulties caused by the number of pages dedicated to the same root.

Regarding issues related to morphology, which influences the root arrangement in dictionaries, it should be noted that, although derivation is productive in verbal morphology, it is not in nominal morphology, with some exceptions. Murcia and Zenia (2015: LXV), for example, offer a list of productive derivative

morphemes in a grammatical summary included in the introduction of their work to facilitate use of the dictionary. However, in the previous section we have offered an example related to nominal derivation and the hesitation showed by dictionaries when it comes to including a supposed consonantal derivative morpheme at the word's root. Some words containing four consonants are collected under GNZR, γ NZR, or XNZR roots, although they seem to be clearly related to NZR triconsonantal root, which has “nose” as basic semantic feature. The meanings of derivative words also have a semantic feature related to “nose” motivated by metaphor or metonymy processes: “bleeding from the nose,” “nostril,” “face, cheek; pejorative,” “cold, flu,” etc.

In the Berber language there is a phenomenon that some authors call “manner derivation” or “expressive derivation” (Chaker, 1973, 1997). Chaker (1995) and Naït-Zerrad (2002) define this phenomenon as a morphological process by affixing derivative morphemes which results in the creation of affective, pejorative, diminutive, or augmentative nouns. It is a frequent phenomenon in body part names. This morphological process consists of the insertion of consonantal sounds at the beginning of the words (the most common of which are [x, q, k, g]), in the repetition of a root consonant or in the reduplication of the whole root. Other authors such as Múrcia (pers. comm, 2014) claim that the phenomena that characterize these words are part of the root, and Galand (1988) does not take a clear position, although he highlights some phenomena typical of diachronic evolution in those words that hinder the recognition of such words as derivatives (Ferrerós, 2014, 2015).

The analyzed case in our study could be included in this type of derivation: the first consonantal sound is typical of expressive derivation and, in fact, one of the listed meanings is pejorative (Serhoual, 2002). Although they seem to be derivative words due to the root consonants, because of the kind of (supposed) affix and because of the semantic relations, all the authors have compiled them under separate entries. In addition, if we observe the information in each entry, we detect a certain hesitation. Thus, for example, Taïfi (1991) places the word *gunzer* “bleeding from the nose” under NZR root, also placed under GNZR.

One might consider that an option that would address the difficulty in arranging derivative words would be to compile them twice as Taïfi (1991) does: once under the simple word entry and another in a separate entry with the derivative morpheme included in the root. However, this decision is not always maintained. If we observe the words collected under the NSR root by Serhoual, some of them are derivatives. In contrast, the ones included under γ NZR do not appear. This lack of coherence occurs in other cases: Serhoual (2002) places *agnic* “lips, pejorative” under the GNC and NC roots yet he places *axnfar* “face, pejorative” under the XNFR root but not under the NFR root.

Root arrangement involves a final problem that we have pointed out in the Introduction related to semantic issues. It should be noted that the organization of meanings inside the entries is generally an issue that can

be improved in all dictionaries, regardless of which languages they describe. As Ibarretxe-Antuñano (2010), an author with a background in cognitivism, claims, in many dictionaries the organization

not only makes it difficult for the reader to grasp the semantic nuisances of these meanings but also hides the real motivated, structured and contextual relationship that exist among these meanings. (Ibarretxe-Antuñano, 2010: 1)

Ibarretxe-Antuñano (2010) proposes the use in the lexicography field of a methodological tool from cognitivism, radial networks. Networks allow the systematic analysis of all the meanings associated with a specific word. In a polysemic word, meanings are motivated by metaphor and metonymy processes and the use of a radial network allows us to determine the type of phenomenon (metaphoric or metonymic) that operates in the semantic extension and it allows us to determine from what meaning another meaning related to the same word derives.

A root arrangement in a dictionary poses a major challenge to the semantic description and meaning structure inside the entries. Moreover, order must be guided by grammatical issues. For example, Bonfour et al. (1995) describe the order of Dallet (1982) entries as follows:

Prenons l'exemple du verbe. On le présente d'abord sous sa forme simple puis dérivée. A l'intérieur de chaque forme, on présente l'impératif de l'aoriste, puis l'intensif et enfin le prétérit. Le nom est présenté aussi avec le souci de rendre compte des variations morphophonologiques. (Bonfour et al., 1995)

An order guided by grammatical features could be compromised if authors try to address linearity in the semantic organization problem and indicate meaning relationships among the words compiled in the same entry. One can observe what happens with NZR meanings compiled by Taïfi (1991). The first word in that entry is a verb meaning "bleeding from the nose." He is respecting grammatical organization. But we must bear in mind that "bleeding from the nose" is a metonymic meaning derived from the prototypical meaning "nose" which does not appear until halfway through the entry.

We do not delve deeper into this issue, but we found it necessary to refer to this matter because we consider that consonantal root arrangement poses a major challenge in semantic description in addition to all the problems mentioned above. We will examine those issues in greater depth in future research. However, this issue and the issue that we have introduced in this paper are matters that should be taken into account in lexicography in order to address some of the remaining problems in Berber lexicography.

Conclusions

In this study, we have presented results related to the two goals listed in the Introduction. First, we gave a small view of the state of Berber lexicography from the 80s to the present day. Second, we analyzed data related to morphology in order to detect and describe certain problems related to the arrangement of words

by consonantal roots. The hypothesis we started with, related to this second goal, was that this type of arrangement is not entirely adequate in Berber lexicography, at least compared to other Afro-Asiatic languages such as Arabic.

With regard to the first purpose, we have presented some characteristics from five recent Berber dictionaries that describe different geographical varieties of the language. These characteristics can be distributed in two groups. First, those that show how Berber lexicography has advanced compared to previous periods in which dictionaries had many theoretical and methodological shortcomings. We have observed how all dictionaries systematically described certain information regarding notation, transcription, grammar and article organization. But there is a second set of features that shows that there is still more progress to be made regarding some aspects, many of which are related to the current situation of the Berber language, in the standardization process: there is no agreement in the use of certain spellings or in the type of transcription (phonetical or phonological). This last issue can also influence the root detection of some words. There is also no agreement in the use of the Latin or Tifinagh alphabets. One of the most significant shortcomings is the lack of monolingual dictionaries.

A characteristic that all dictionaries show, and regarding which some progress remains to be made, is the one related to the second goal: lexical arrangement by consonantal roots. In fact, all the described dictionaries add a justification text in order to explain why the authors decided to use this type of arrangement. This single fact already indicates that this arrangement involves certain problems: “c’est, sans doute, dans ce secteur et celui du classement par racine [...] que des progrès restent à faire” (Bonfour et al. 1995). Although some years have passed, it seems that progress must be continued in this regard.

We have seen that much of the vocabulary is not morphologically analyzable (especially for diachronic causes) by non-linguistics users, that there are many homophonic roots, and many supposed derivative morphemes, especially in nominal lexical items, are not productive. All of these issues make root detection challenging, especially with regards to non-specialized speakers. However, some other morphological features and traditions in the use of this type of arrangement make change difficult. Nevertheless, we believe that describing and taking into account all of these issues can make lexicographical research advance in order to improve word arrangements in future works. Finally, we have taken into account that root arrangement involves an added challenge in terms of the semantic description of the words contained therein. It is for this reason that the use of a methodological tool such as cognitivism’s radial networks, which allow a systematic analysis of meanings, can address the fact that dictionaries often compile meanings in a linear manner and without indications about semantic relationships.

References

BOUKHRIS, F., BOUMALK, A., EL MOUJAHID, E. and SOUIFI, H. (2008): *La nouvelle grammaire de l’amazighe*, Rabat, Publications de l’Institut Royal de la Culture Amazighe.

- BOUNFOUR, A, LANFRY, J. and CHAKER, S.. (1995): «Dictionnaires Berbères », in *Encyclopedie berbere*. Vol. XV, Aix-en- Provence, Edisud.
- CHAKER, S. (1973): « Derivés de manière en berbère», en *Comptes rendus du Groupe Linguistique d'Études Chamito-sémitiques (GLECS)*, vol. XVII, Paris, Geuthner.
- CHAKER, S. (1995): «Dérivation», in *Encyclopedie berbere*. Vol. XV, Aix-en- Provence, Edisud.
- CHAKER, S. (1997): «Expréssivité», in *Encyclopedie berbere*. Vol. XVIII, Aix-en- Provence, Edisud.
- CHAKER, S. (2013) : « La racine en berbère : réalité synchronique ou diachronique ? Implications lexicographiques », in *Linguistique arabe et hamito-sémitique* (seminar, 13/11/2013), Aix-en-Provence.
- DALLET, J. M. (1982) : *Dictionnaire Kabyle-Français. Parler des At Mangellat, Algérie*, Paris, Societé d'études linguistiques et anthropologiques de France.
- Dallet, J. M. (1985) : *Dictionnaire Français-Kabyle. Parler des At Mangellat, Algérie*, Paris, Societé d'études linguistiques et anthropologiques de France.
- El Aissati, A. (2014) : « Script choice and power struggle in Morocco», in Juffermans, K et al. (eds.) *African Literacies: Ideologies, Scripts, Education*, Newcastle upon Tyne, Cambridge Scholars Publishing.
- FAUCAULD, C. (1951-1952) : *Dictionnaire Touareg-Français. Dialecte de l'Ahaggar*, Paris, Imprimerie Nationale de France.
- FERRERÓS, C. (2014): «Sustantivos que designan partes del cuerpo en rifeño: una reflexión sobre la ordenación léxica en los diccionarios bereberes», in CAMUS, B. Morfología y diccionarios. Anexos de Revista de Lexicografía. Universidade da Coruña, Serivzo de Publicacións.
- FERRERÓS, C. (2015): *Categorització semàntica de les parts del cos en català i en amazic: un estudi comparatiu*. Tesi doctoral, Universitat de Girona.
- GALAND, L. (1988): «Le berbère », en Perrot, J. (dir.) *Les langues dans le monde ancien et moderne: Langues Chamito-Sémitiques*, Paris, Éditions du CNRS.
- IBARRETXE-ANTUÑANO, I. (2010). «Lexicografía y lingüística cognitiva» en Revista española de lingüística aplicada (23), 195-214.
- LAMUELA, X. (2003): *El berber: estudi comparatiu entre la gramàtica del català i la del berber o amazig*, Barcelona, Generalitat de Catalunya, Departament de Benestar i Família.
- MÚRCIA, C. and ZENIA, S. (2015). *Diccionari català-amazic / amazic-català (estàndard del diasistema amazic septentrional)*. Llibres de l'índex, Barcelona.
- MÚRCIA, C. (2011): *La llengua amaziga a l'antiguitat a partir de les fonts gregues i llatines*. Barcelona, PPU. Promocions i Publicacions Universitàries. Colección Cum Laude, 4.

NAÏT-ZERRAD, K. (1997, 1999, 2002): *Dictionnaire des racines berbères (formes attestées)* (tres vol., hasta la <g>: vol. I, 1997; vol II, 1999; vol. III, 2002), Paris / Lovain, Éditions Peeters.

NAÏT-ZERRAD, K. (2002): «Les préfixes expressifs en berbère », in NAÏT-ZERRAD, K (ed.) *Articles de linguistique berbère. Méorial Werner Vycichl*, Paris, L'Harmattan.

SERHOUAL, M. (2002): *Dictionnaire tarifit-français*. Thèse de doctorat d'Etat des lettres. Tétouan, Université Abdelmalek Essaâdi, Faculté des Lettres et des Sciences Humaines.

TAIFI, M. (1991): *Dictionnaire Tamazight-Français (parlers du Maroc central)*. Paris, L'Harmattan.

ON LEXICAL EQUIVALENCE IN THE BILINGUAL DICTIONARY PHILOSOPHICAL AND MENTAL-REPRESENTATION REFLECTIONS

Cuilian Zhao

Fudan University, China

Abstract

The thesis explores how two philosophical themes (the indeterminacy of translation and linguistic relativism) shed light on lexical equivalence in the bilingual dictionary from the perspective of mental lexical representation. After a brief delimitation of bilingual dictionary equivalence, the present writer introduces the two relevant themes of linguistic philosophy and two classic models of bilingual mental lexical representation. Based on such a theoretical framework, an operational definition of equivalence is proposed: equivalence between the source language and the target language is considered at three levels – lexical, semantic, and world knowledge levels; equivalence depends on the degree of closeness from the formal (lexical) level to the conceptual (semantic + world knowledge) levels, i.e. the degree of lexical closeness (or transparency), the proportion of shared semantic features, and the proportion of shared world knowledge in the mental lexicons of language users in their respective linguistic environments. Two pairs of English-Chinese expressions (an abstract pair *terrorism* – 恐怖主义 and a concrete pair *piano* – 钢琴) are selected for in-depth analysis in order to show their degrees of (non-)equivalence. The paper concludes by pointing out that, first, both abstract and concrete lexical pairs have non-shared parts in the mental representations in their respective cultures; second, bilingual dictionary translation influences the thinking patterns of dictionary users; third, total equivalence of dictionary translation can be approached, if not attained.

Key Words: bilingual lexical equivalence, mental representation, indeterminacy of translation, linguistic relativity

Introduction

Does *terrorism* mean 恐怖主义 (terror-ism: a set of ideas or a system of beliefs in the use of violent action)?

Is *vested interest* equivalent to 既得利益 (secured benefit)? What kind of mental representations do the expressions *piano* and 钢琴 (steel(-string) instrument) respectively produce apart from serving as labels (in different languages) appended to the same musical instrument? Is the acronym 双规 (double-designate) equivalent in form to its possible translation “to be subject to investigation at a designated time and place”? These words or expressions have been accepted as pairs of translation equivalents, but it is questionable whether such pairs are equivalent from form to meaning, and whether they can ensure understanding or evoke identical cultural connections in the (bilingual) dictionary-user’s mental lexicon. It is therefore advisable that the bilingual dictionary compiler or lexicographer delve deeper into the matter of equivalence.

Equivalence is one of the central concerns of bilingual lexicographers. According to Zgusta, “The bilingual lexicographer’s most important duty is to find in the target language such lexical units as are equivalent to the lexical units of the source language.” (Zgusta 1971: 312) Ideally, equivalence is realized by way of providing words or expressions that refer to similar entities in two cultures. However, for two languages with divergent social and cultural backgrounds, translational equivalence is often difficult to establish. Since the words or expressions are usually closely connected with different cultures, absolutely equivalent words from two languages are rare. On the contrary, it is often the case that words in one language have no equivalents in another, which results in linguistic or cultural lacunarity. It is generally accepted that the bilingual dictionary serves the role of bridging interlingual or cross-cultural gaps. (e.g. Zgusta 1971: 312; Manning 1990: 159; Szerszunowicz 2015) In addition, by patching up the linguistic or cultural lacunarity, the bilingual dictionary might introduce the user to novel concepts from a different culture, thus exerting a potential influence on the user’s mind. The questions that naturally arise are, if total equivalence is hard to establish, how can the bilingual dictionary ensure effective transfer of information or knowledge between cultures? Can lexical units treated in the bilingual dictionary exert similar influence on the bilingual dictionary user’s mind as it does on the native speaker? Is there any principled approach? Theoretical framework?

The above questions are essentially questions about cross-linguistic and cross-cultural transfer of information, and about the representation of bilingual mental lexicon. To answer these questions, we turn to two related themes – the indeterminacy of translation and linguistic relativity, and explore how these can shed light on the issue of bilingual lexicographic equivalence from the perspective of mental lexicon representation.

Method

In what follows, after a brief delimitation of equivalence in bilingual lexicography, this paper discusses Quine's indeterminacy of translation (plus his inscrutability of reference) and Whorf's linguistic relativity hypothesis, and then introduces two classic models of bilingual mental lexical representation. Within this theoretical framework, two pairs of words (an abstract pair *terrorism*-恐怖主义, and a concrete pair *piano*-钢琴) are selected for analysis at various levels. The purpose is to find out the degree of equivalence of the word pairs.

1. Equivalence in the bilingual dictionary

This section sets out to clarify the notion of lexicographic equivalence. We first review lexicographic definitions, drawing insights from relevant philosophical themes and psycholinguistic models. Finally we provide an operational definition of lexicographic equivalence.

Zgusta explains *equivalent* as "such a lexical unit of the target language which has the same lexical meaning as the respective lexical unit of the source language" (Zgusta 1971: 312), which entails that bilingual lexicographic equivalence is essentially lexical semantic equivalence. Some distance away in the same chapter, he further distinguishes two categories: *translational (insertable)* and *explanatory (descriptive)* equivalents (Zgusta 1971: 319), which expands the considerations of equivalence from the lexical level to the sentential or contextual level, and from a generalized, abstract, and stable static relationship to a concrete, imagery, and dynamic relationship.

Hartman & James (1998: 51) define *equivalence* as: "The relationship between words or phrases ... which share the same MEANING. Because of the problem of ANISOMORPHISM, equivalence is 'partial' or 'relative' rather than 'full' or 'exact' ...," "equivalence implies interlingual correspondence of DESIGNATIONS for identical CONCEPTS." For them, an *equivalent* is "A word or phrase in one language which corresponds in MEANING to a word or phrase in another language." In Hartman and James' definitions of *equivalence* and *equivalent*, we find interlingual corresponding entities ("words or phrases"), focus of comparison ("meaning"), and relationship between the entities ("identical concepts", "corresponds in meaning"). They also point out the partiality, relativity, and asymmetry of this correspondence.

The above definitions indicate that lexicographic equivalence involves three aspects: two entities, their relationship, and the evaluation of this relationship. The idea is that there exist in the source language and the target language two corresponding entities, which are related by degrees (e.g. partial versus full equivalence) in different respects (e.g. semantic, conceptual, descriptive, translational, functional). But such considerations are problematic in that, firstly, it is doubtful whether such entities exist (indeterminacy of reference, or ontological relativity); secondly, if there are such corresponding entities, to what extent they

are equivalent (indeterminacy of translation); thirdly, if there is no equivalent entity in the target language, whether or not such an entity should be “created” in the bilingual dictionary to facilitate communication. These questions verge on Quine’s doctrine of the indeterminacy of translation (and the inscrutability of reference).

2. Quine’s indeterminacy of translation

In view of whether an utterance (word or phrase) has a definite meaning and whether a term refers to a single object, Quine (1960) proposed his famous *indetermination of translation* and *inscrutability of reference*.

Quine’s argument for his indeterminacy thesis is that mutually incompatible translation manuals can be made equally compatible with all the possible evidence by compensatorily juggling the translation of the apparatus of individuation. In his famous thought experiment of radical translation, a field linguist sets about translating a hitherto unknown language that has no historical or cultural connections with any known language. The total empirical data available to the linguist consist of the observable behaviour of native speakers amid publicly observable circumstances:

... manuals for translating one language into another can be set up in divergent ways, all compatible with the totality of speech dispositions, yet incompatible with one another. (Quine 1960: 27)

Despite the idealized context, the linguist’s completed manual for translating the foreign language into the linguist’s home language is underdetermined by all of the possible empirical data. Specifically, the translation of the terms and meanings of theoretical sentences are underdetermined. Quine’s conclusion from this thought experiment is that the translation of theoretical sentences is not merely underdetermined but indeterminate. The core idea is that the same foreign sentence can be translated equally well by two or more different home language sentences.

Closely related to the indeterminacy of theoretical sentences is the indeterminacy of reference or inscrutability of reference. The inscrutability of reference is exactly the same as the indeterminacy of translation, verified by the same evidence, except that it applies to terms rather than whole sentences (Quine 1968). The point here is that stimulus meaning does not determine reference, that we cannot know the exact reference of a term.

... we could equate a native expression with any of the disparate English terms ‘rabbit’, ‘rabbit stage’, ‘undetached rabbit part’, etc., and still, by compensatorily juggling the translation of numerical identity and associated particles, preserve conformity to stimulus meanings of occasion sentences. (Quine 1960: 54)

Quine also seems to suggest that if there is indeterminacy of translation at all, it holds in the “domestic” case: reference is after all *le*. Quine agrees with Dewey that there is no “private language” (Quine 1968:

199-200). Because of the non-existence of private language, we have no exclusive right to our personal language: if the inscrutability of reference applies to other people, it applies to ourselves (the first person) as well: we cannot know the reference of a term used by ourselves. Some scholars complain that this idea is counterintuitive and distorts common sense (e.g. Searle 1987).

What does Quine's indeterminacy theory imply for inter-lingual translation? In the first place, "indeterminacy" does not mean that translation is impossible. Rather, it means that translation is too easy, with too many right answers. Beyond the parameters set by linguistic behaviour, there is nothing to be wrong or right about. Therefore, in translating between relatively remote languages and cultures, inequivalent sentences of one language will often do equally well as rough translations of a single sentence of another. And "there is not even ... an objective matter of fact to be right or wrong about" (Quine 1960: 73). In general, the effect of two languages having different resources will be that there is no exact translation between them, whether we are talking about the translation of whole sentences or of referring expression.

When it comes to bilingual lexicographic equivalence, Quine's indeterminacy theory has its explanative power. Firstly, lexicographic equivalence is a matter of degree: completely equivalent entities do not exist (indeterminacy of reference, or ontological relativity). Secondly, lexicographic translation involves language as a hole. As Quine likes to put it, the indeterminacy of translation "cuts across extension and intension alike" (Quine 1968). Also, as a variety of holism, Wittgenstein stated: "To understand a sentence means to understand a language" (Wittgenstein 2009: 87). Indeterminacy of translation means the many possibilities of choice-making in the overall picture of the dictionary or the language when it comes to corresponding entities in bilingual lexicography. Therefore, our concern can be so modified as to the extent of translational equivalence in bilingual lexicography, and the degree of our confidence about the reference of an expression. Specifically, to what extent can translational equivalence be established in the bilingual dictionary? This will be explored in Section 4 when discussing the bilingual mental lexical representation.

3. Whorf's linguistic relativity

Since the bilingual dictionary can play the dual role of helping with both decoding and encoding activities, it is advisable that the bilingual lexicographer take into account the user's dual cognitive needs of understanding and using the language. Below we explore Whorf's linguistic relativity hypothesis that can shed light on the bilingual dictionary user's cognitive needs.

Literature distinguishes two Whorfian hypothesis: linguistic determinism and linguistic relativity (e.g. Carroll 2008: 396; Penn 1972: 15). The former is the strong form, referring to the notion that language determines certain nonlinguistic cognitive processes; the latter is the weaker form, claiming that the cognitive processes determined are different from language to language. Therefore, it can be said that the thinking patterns of speakers in different languages are different. In his quote about the identity of language

and thought (Carroll 2008: 396), he made explicit three notions: first, languages ‘‘cut up’’ reality in different ways; second, these language differences are covert or unconscious; third, these language differences influence our worldview.

Whorf provided lexical and grammatical examples to show that linguistic determinism and relativity are valid concepts. For example, in an American Indian language of Hopi, one word covers all flying objects (such as flying insects, airplanes, aviators) except birds, which is too broad a category for most people in other parts of the world; but for the Eskimos, using one word *snow* to refer to various forms of snow (falling snow, snow on the ground, snow packed hard like ice, slushy snow, wind-driven flying snow) is all-inclusive and unthinkable. Whorf believes that there is no ‘‘natural’’ way to ‘‘carve up’’ reality, and different languages dissect reality in quite different ways.

For Whorf, these different carvings of reality lead to different thinking patterns. That is, when we frequently encounter a word, it may influence our habitual thinking pattern. His ‘‘empty gasoline drum’’ example serves to show the power of words influencing the thinking process.

The Whorf hypothesis is an important theory about the relationship between language, culture and thinking. Its strong and weak forms explain to varying degrees that differences in words and grammar in different cultures influence the thinking patterns of the language users. In view of bilingual lexicographic translation, the Whorf hypothesis suggests that, once a translation makes its way into the dictionary entry, it will most probably be regarded as a correct or model way of rendering by the dictionary user. In time the user will gradually register in his mind the translation as an equivalent, which will gradually influence their way of thinking.

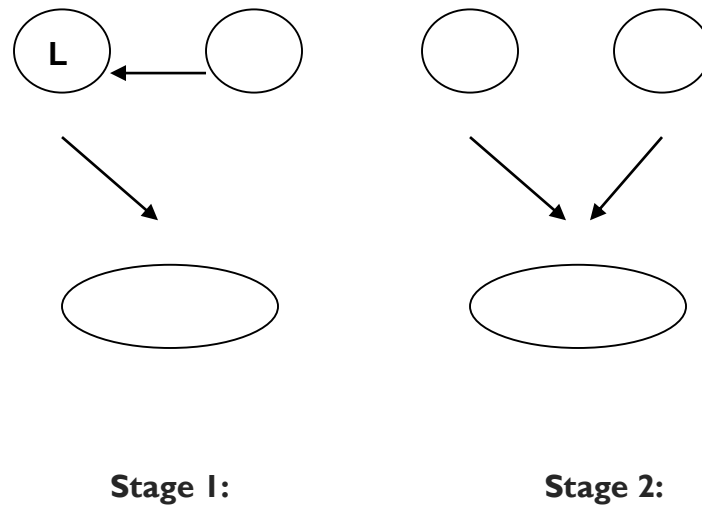
When it comes to the influence of lexicographic translation on the user’s thinking pattern, we are well on the way to exploring the bilingual mental lexicon. In what follows, we will first introduce the mental representation of the bilingual lexicon, and then, evoking what is left unsaid by the end of Section 2, we will set out to explore the issue of bilingual lexicographic equivalence with concrete examples.

4. Mental lexical representation and bilingual lexicographic equivalence

4.1 Two classic models of the L2 learner’s mental lexical representation

From the perspective of psycholinguistics, there are at least two levels of lexical representation in the mental lexicon, i.e. the lexical level (incorporating the lemma and lexeme levels) and the conceptual level. According to de Groot et al. (1995), when the L2 learner acquires a new L2 word, the connection between the L2 lexical node and its L1 equivalent lexical node is strong, i.e., there is direct connection between the L1 and L2 lexical levels. The learner understands the L2 word by way of activating the relative concept through the L1 word. This is referred to as ‘‘lexical connection.’’ With the development of the L2 learner’s proficiency, the connection between the L2 word form and the relative concept in the mental lexicon

becomes stronger, and the lexical connection gradually diminishes. Finally, the L2 learner can directly activate the relative concept through the L2 word. This is referred to as “conceptual mediation.” The development of the L2 mental lexicon is illustrated in Figure 1 adapted from de Groot et al. (1995), with amendments. (The arrows indicate connections and comprehension processes.)



This hypothesis shows that the learner’s mental representation for a newly acquired L2 word is biased towards lexical connection, but the mental representation for a familiar L2 word is biased towards conceptual connection. Here the “conceptual” representation refers to the L2 learner’s conceptual representation. But the question is, in the L2 learner’s conceptual representation, are the semantic components of the source word all represented? Is the world knowledge associated with the source word represented? Therefore, it is necessary to analyze the semantic components of the source word, and to divide the conceptual representation into semantic representation and knowledge representation.

De Groot and his colleagues (de Groot 1992; Van Hell & de Groot 1998) proposed the “distributed feature” model of bilingual lexical representation (Figure 2). In this model, word meaning is represented in memory as a set of semantic features, some of which are shared between a pair of translations, whereas others are unique to either the L1 word or the L2 word. Translations of concrete words and cognates share more of these semantic features than translations of abstract words and non-cognates.

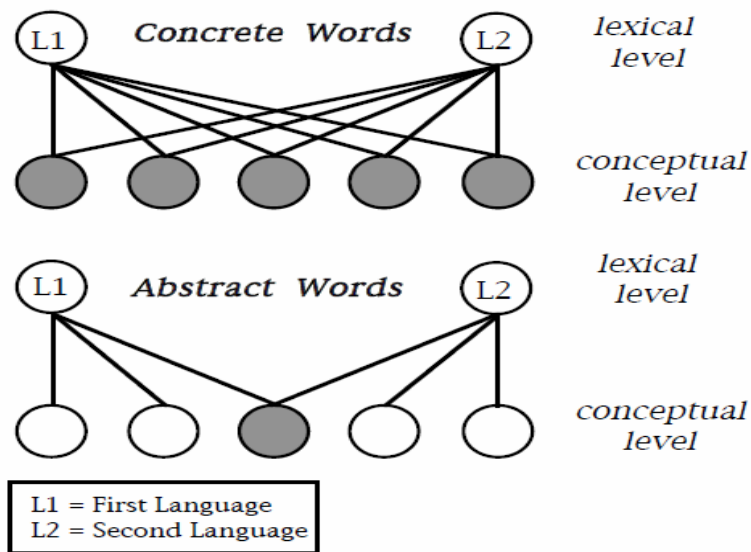


Figure 2. Distributed feature model of bilingual lexical representation

The distributed feature model shows that the two languages of the bilingual share the semantic system, and both languages can access these features that constitute the semantic elements. Which of a word's many meanings should be assigned to it when it is encountered in speech or reading depends on the context of use. The meanings of an abstract word is more context-dependent: if the context of a word undergoing the mental lexicon processing shifts with its language and culture, then the meaning of an abstract word depends more on context than a concrete word.

The distributed feature model partially answers our question. That is, in the L2 learner's conceptual representation, the semantic components of the source word cannot not fully represented. By analogy, beyond the semantic components, the world knowledge relevant to the source word, be it concrete or abstract, will not be fully represented as well.

4.2 Bilingual lexicographic equivalence: an operational definition

Based on the above assumptions and drawing insights from philosophical reflections, we further delimitate cross-language equivalence, and propose an operational definition. We presume that mental representation of the dictionary user locating the source-word meaning is biased towards lexical-connection. The bilingual dictionary is to help the dictionary user establish this connection. The bilingual dictionary translates (or interprets) the source word and yields a translation. Equivalence between the source language and the target language manifests itself at three levels: lexical, semantic, and world knowledge levels; the degree of equivalence depends on the degree of closeness from the formal (lexical) to the conceptual (semantic +

world knowledge) levels, i.e. the degree of lexical closeness (or transparency), the proportion of shared semantic features, and the proportion of shared world knowledge in the mental lexicons of language users in their respective linguistic environments.

5. Case study

In this section, we look into bilingual lexicographic equivalence from the perspective of mental lexical representation. Two pairs of words are selected for analysis: an abstract pair (terrorism-恐怖主义) and a concrete pair (piano-钢琴). The abstract pair is representative of matching two existing entities from two languages; the concrete pair is representative of creating a previously non-existent entity in the translated language.

5.1 *terrorism* – 恐怖主义

Terrorism and 恐怖主义 are two concepts from English and Chinese respectively. The current practice of bilingual lexicography is to treat the two as equivalents. The one-to-one mapping itself is questionable. For, according to Quine's indeterminacy of translation, either *terrorism* or 恐怖主义 is open to interpretations. So 恐怖主义 is not necessarily the one correct translation of *terrorism*, nor vice versa. What's more, before rendering *terrorism* and 恐怖主义 equivalent, we have to locate the possible entities these two terms refer to respectively. Within the theoretical framework of the indeterminacy theory and L2 mental lexical representation models, we shall assess the degree of equivalence of the word pair.

We shall set out to locate the possible referent of *terrorism*. According to Wikipedia (<https://en.wikipedia.org/wiki/Terrorism>): “there are over 109 different definitions of terrorism;” “the international community has been slow to formulate a universally agreed, legally binding definition of this crime;” “these difficulties arise from the fact that the term ‘terrorism’ is politically and emotionally charged.” These explanations seem to verge on Quine's inscrutability of reference: the referent of *terrorism* seems elusive; it is hard to pin down a generally accepted definition.

Let's look at the dictionary definition of this concept. The relevant terms and definitions in the *Oxford Advanced Learner's English-Chinese Dictionary* (9th ed.) (hereafter referred to as *OALECD*) are as follows:

terror ... **3** violent action or the threat of violent action that is intended to cause fear, usually for political purposes ... **SYN** TERRORISM

terrorism the use of violent action in order to achieve political aims or to force a government to act

In this dictionary, the definition of *terrorism* and the third sense of *terror* are synonymous, both referring to “(threat of) violent action (for political purposes)”, which is confirmed by the synonym label **SYN** **TERRORISM** at the end of sense 3 under the entry *terror*. Therefore, in the *OECD*, the core meaning of *terrorism* is “(the use of) violent action”.

The online dictionary provided three definitions of *terrorism*:

1. the use of violence and threats to intimidate or coerce, esp. for political purposes.
2. the state of fear and submission produced by terrorism or terrorization.
3. a terroristic method of governing or of resisting a government.

<http://dictionary.reference.com/browse/terrorism>

The first definition is similar to the *OECD* definition: “the use of violent action and threats.” Both have the key element “action”. The second definition refers to “the state of fear ...,” and the third refers to “terroristic method.”

Wikipedia’s interpretation of this concept is:

Terrorism is the systematic use of terror, often violent, especially as a means of coercion.

<http://en.wikipedia.org/wiki/Terrorism>

Here again, the key expression is “use of terror.” *Terrorism* is an act.

The above sources of information differ in the design and definition of senses, but from these we glean two points: 1) *terrorism* is an “action” or “use of violence”; 2) *terrorism* is a polysemous concept.

“Polysemy” partly echoes Quine’s “inscrutability of reference”: when detached from context, we cannot know for sure what *terrorism* refers to.

Again let’s look at “action” or “use of violence.” The word *terrorism* consists of two morphemes: *terror* and *-ism*. The meaning of *terror* here is unambiguous: “(threat of) violent action.” However, the morpheme *-ism* is ambiguous. In the *OECD*, *-ism* has 6 senses:

-ism **1** the action or result of ...的行为 (或结果) : *criticism* 批评 **2** the state or quality of ...的状态 (或品质) : *heroism* 英勇 **3** the teaching, system or movement of ...的教义 (或体系、运动) : *Buddhism* 佛教 **4** unfair treatment or hatred for the reason mentioned 因...的不公平对待 (或敌意) :

racism 种族偏见 **5** a feature of language of the type mentioned ...的语言特点 : *Americanism* 美国英语

的特点 ◇ *colloquialism* 口语体 **6** a medical condition or disease 健康状况 ; 疾病 : *alcoholism* 酒精中

毒

The above division of senses indicates that, 1) *-ism* is polysemous; 2) its most frequently used sense (the 1st sense) is “the action or result of”; 3) the 3rd sense is related to the 1st. In Table 1, the senses of *-ism* in the *OECD* are aligned with the definitions of *terrorism* alluded to in the preceding part of this section.

Table 1: Contrasting the definitions of *terrorism* and *-ism*

	<i>terrorism</i>	<i>-ism</i>
<i>OALECD</i>	the use of violent action in order to achieve political aims or to force a government to act	1 the action or result of
dictionary.reference.com	1. the use of violence and threats to intimidate or coerce, esp. for political purposes.	1 the action or result of
	2. the state of fear and submission produced by terrorism or terrorization.	2 the state or quality of
	3. a terroristic method of governing or of resisting a government	?
Wikipedia	the systematic use of terror, often violent, especially as a means of coercion	1 the action or result of 3 the teaching, system or movement of

Table 1 shows that, 1) just like the referential inscrutability of *-ism*, the reference of *terrorism* is indeterminate, that is, it is polysemous; 2) the key sense of *terrorism* is aligned with the most frequently used sense of *-ism*, focusing on “action”; 3) the semantic component associated with the 3rd sense of *-ism*, i.e. “the teaching, system or movement of,” is found in Wikipedia’s definition, i.e. “systematic use of terror,” where “systematic” modifies the central noun “use.”

Let us look at the definition of 恐怖主义. The *Modern Chinese Dictionary* (6th Edition) (hereinafter referred to as the *MCD*) provides the following definition:

【恐怖主义】蓄意通过暴力手段（如制造爆炸事件、劫持飞机、绑架等），造成平民或非战斗人员伤亡与财产损失，以达到某种政治目的的行为和主张。

{[terror-ism] an act or position to achieve certain political goals, through the deliberate use of violent means (such as bombing, hijacking aircraft, kidnapping, etc.), causing civilian or non-combatant casualties and property losses }

The definition contains two central nouns: 行为(act) and 主张(position). Therefore, 恐怖主义 can be illustrated with the following equation:

恐怖主义 = 恐怖行为 + 以暴力手段达到政治目的的主张

{terror-ism = act of terror + position advocating the use of violence to achieve political goals }

This equation echoes the lexical components of the expression 恐怖主义, namely:

恐怖主义 = 恐怖 + 主义 = 恐怖行为 + 以暴力手段达到政治目的的主张

{terror-ism = terror + -ism (doctrine) = act of terror + position advocating the use of violence to achieve political goals }

According to this equation, the meaning of 恐怖主义 is the sum of the meanings of its lexical components.

The question to ask here is, what constitute the respective referents of the components 恐怖 and 主义? The *Greater Chinese Dictionary* (hereafter referred to as the *GCD*) explains the two lexical components as follows:

【恐怖】1感到可怕的畏惧。亦谓令人畏惧。2威胁；恫吓。

{[terror] 1 intimidating fear; (also) to frighten 2 threat; intimidation }

【主义】 1谨守仁义。2对事情的主张。3犹主旨，主体。4形成系统的理论学说或思想体系。马克思主义，达尔文主义。5一定的社会制度或政治经济体系。社会主义；资本主义。6思想作风。自由主义；主观主义。

{[doctrine] 1 commitment to benevolence and righteousness 2 assertion; position; opinion 3 theme; dominant idea; gist 4 theoretical doctrine or ideology forming a system: Marxism, Darwinism 5 social system or political and economic system: socialism; capitalism 6 ideological style: liberalism; subjectivism}

In Table 2, the relevant senses of 恐怖 and 主义 in the *GCD* are aligned with the definitions of 恐怖主义 alluded to in the preceding part of this section.

Table 2: Contrasting 恐怖主义 and the definitions of its lexical components

恐怖主义 (terror-ism)	恐怖 (terror)	主义 (-ism)
恐怖行为 + 以暴力手段达到政治目的的主张 {terror-ism = act of terror + position advocating the use of violence to achieve political goals}	威胁；恫吓 {threat; intimidation}	对事情的主张 {assertion; position; opinion}

However, according to the structural analysis, 恐怖主义 is a modifier-modified phrase, 恐怖(terror) the modifier, 主义 (-ism) the modified (and thus the central noun). Therefore, the core meaning of 恐怖主义 is “-ism,” a kind of “position.”

In Table 3, the definition of 恐怖主义 is aligned with the analysis of *terrorism* alluded to in the preceding part of this section. The contrast shows that *terrorism* and 恐怖主义 are partially equivalent.

Table 3: Contrasting the definitions of *terrorism* and 恐怖主义

	<i>terrorism</i>	恐怖主义(terror-ism)
definition	act of terror; use of violent action	恐怖行为 + 以暴力手段达到政治目的的主张 (act of terror + position advocating the use of violence to achieve political goals)
central noun	Act	主张 (position)

By this point we can assess the degree of equivalence of *terrorism* and 恐怖主义. According to the operational definition of equivalence proposed in Section 4.2, the degree equivalence depends on the degree of closeness from the formal (lexical) to the conceptual (semantic + world knowledge) levels (Table 4).

Table 4: Degree of equivalence of *terrorism* and 恐怖主义

Level		<i>terrorism</i>	恐怖主义 (terror-ism)	degree of closeness	
Lexical	modifier	terror	恐怖 (terror)	high	low
	Modified/Central morpheme	-ism (action)	主义 (-ism)	low	
semantic	Semantic component	act of terror	恐怖行为 (act if terror)	high	low
		use of violent action	以暴力手段达到政治目的的主张 (position advocating the use of violence to achieve political goals)	low	
	Central noun	act; use	主张 (position)	low	
World knowledge		illegal act of violence:	组织、制度、政治目的 :		

	hijacking; bombing; kidnapping; “9·11”	基地组织；本·拉登； “9·11”事件 (organized, systemic, with political purposes: Al Qaeda; bin Laden; “9·11”)	low
--	---	---	-----

Table 4 shows that the degrees of closeness between *terrorism* and 恐怖主义 verge on the low side at all levels. That is, this word pair is not equivalent at the lexical, semantic, and world knowledge levels. The non-equivalence of *terrorism*—恐怖主义 shows that it is nothing easy to locate “equivalent” entities, especially abstract ones, in two languages. On the one hand, there is the inscrutability of reference for such abstract terms in their domestic languages, which renders them ambiguous when decontextualized. On the other hand, it is generally the case that semantic features of “equivalent” abstract words in two languages are asymmetrically represented in the mental lexicon. Lu Gusun (2012) once likened the process of finding an equivalent to “bridging” and “arriving”:

...[translation is] starting from one language and “arriving” on the opposite bank of another language ... But, more often than not, the two languages can be radically different because of the respective cultures attaching to them. Their respective secrets lie far away on either end of the bridge, buried deeply in the hinterland. It is simply impossible to locate a corresponding point by crossing the bridge itself, let alone “arriving”. (Lu 2012)

Two seemingly similar entities in two languages, especially abstract concepts, may share limited semantic features and differ in their mental representations. Therefore, it is not advisable to simply treat them as equivalents.

In contrast, when there does not exist any similar or equivalent entity in the target language, the usual lexicographic practice is to define the source entity or “create” an entity in the target language. Apparently, providing a novel name is much easier than locating an existing one. Nevertheless, multiple factors must be considered so as to faithfully convey information of the word in the source language. In the next section, by way of analyzing the word pair of *piano*—钢琴, we look into the case of introducing novel entities in bilingual lexicography.

5.2 *piano*—钢琴

Bilingual dictionary users, especially lower level users, rely heavily on translated words or expressions for their understanding of the source words or expressions. On the other hand, according to the Whorf hypothesis, dictionary users (or, on a broader scale, foreign language learners) repeatedly access translated words or expressions during the process of foreign language learning. Over time, the message conveyed by the translated language will subtly influence their ways thinking and even their world view. It is thus suggested that when translating or localizing foreign words, bilingual lexicographers need to be as consistent as possible with the source words, from form to meaning, and at other levels such as imagery and association.

Take the word *piano* as an example. Its translation is 钢琴, referring to the same instrument. But, leaving the material ontology aside, our concern here is whether the two expressions are equivalent from form to content, and whether they evoke the same representation in the dictionary user's mental lexicon.

Let's first look at the information conveyed by the source word *piano*. The relevant entries and definitions in the *OALECD* are as follows:

piano *noun, adv.*

■ *noun* // (*pl.*-os) (also *old-fashioned, formal pianoforte* //) a large musical instrument played by pressing the black and white keys on the keyboard. The sound is produced by small HAMMERS hitting the metal strings inside the piano.

■ *adv.* // (*abbr.* **p**) (*music*) played or sung quietly

forte // *noun, adv.*

■ *noun* [*sing.*] a thing that sb does particularly well

■ *adv.* (*music*) played or sung loudly **opp** PIANO

The *piano* entry yields the following information:

(1) morphological (lexical level): The *old-fashioned* or *formal* script of *piano* is **pianoforte** (piano + forte). In musical expressions, *piano* and *forte* are a pair of antonyms, the former meaning “quietly,” and the latter “loudly.”

(2) definitive (semantic level): large musical instrument; played by pressing the black and white keys; sound produced by small hammers hitting the metal strings.

Thus, the entry word *piano* refers to a musical instrument, but message conveyed by its written form is about musical expression. As an abbreviation for *pianoforte* (meaning “gentle + powerful,” or “soft+loud”), the expression *piano* retains the “gentle” part of the word. As designation, *piano* is not about the physical

composition of the instrument, but rather foregrounds the musical expression: soft key touch generating graceful notes.

The source word *piano* aside, what kind of information does the translation 钢琴 convey? The definition in the *MCD* runs as follows:

【钢琴】gāngqín 名 键盘乐器，内部装有许多钢丝弦和包有绒毡的木槌，一按键就能带动木槌敲打钢丝弦而发出声音。

{[steel-instrument] *noun* keyboard instrument, with many steel strings and cushioned wooden hammers inside; a press on the key triggers the hammer to hit a steel string and produce sound}

This definition briefly describes the mechanism of the piano and how to work it. Let's suppose such is the message picked up by the dictionary user. What is the overall impression on the dictionary user? From the perspective of mental lexical representation, what information is foregrounded?

The definition consists of 40 Chinese characters, from which we glean expressions in three categories: “material”, “operating verb” and “timbre” or “sound quality.” We then dichotomize the “metallicity” or “noise” of these expressions (an [1] means “yes”, and a [0] means “neutral” or “zero expression” (See Table 5).

Table 5: Message conveyed in the definition of 钢琴

	Message	metallicity
Material	键盘[1]、键[1]、钢丝弦2[1]、木槌2[1]、绒毡 1[0] {keyboard; key; steel strings; wood hammer; down cushion}	7 [1] , 1[0]
Operational verb	按[0]、带动[0]、敲打[1]、发出[0] {press; trigger; hit; produce}	1[1] , 3[0]
Timbre	声音[0] (sound)	1[0]

As illustrated in Table 5, the expressions denoting hard materials are “keyboard, keys, wire strings, wooden hammer,” accounting for 87% of the materials. Like the entry expression 钢琴 (literally, steel-instrument), these expressions foreground the “metallic” impression (or noise) in the dictionary user’s mental representation by way of reiterating the “metallic nature” and “hardness” of the instrument.

The verbs for the keyboard include “press, trigger, hit, and produce (sound)”, among which the verbs “press, trigger, and produce” do not carry the “strength” of operating the keys, and can thus be rated as neutral verbs. The verb “hit” is an action with the immediate effect of striking a note, and carries the strength of operating the keys. The message picked up by the dictionary user is a hard “knock.”

The expression “sound” is a neutral word, which does not carry any special timbre.

We do the simple calculation below to obtain the overall “metallic (noise) value” of these expressions in Table 5:

$$(7 + 1) \times [1] + (1 + 3 + 1) \times [0] = 8$$

The result obtained is a positive number 8, which means that the overall impression of 钢琴 on the dictionary user is positively “metallic (noise).” In other words, the “metallicity (noise)” conveyed by 钢琴 is prominently foregrounded in the dictionary user’s mind.

In Table 6, we compare the source word *piano* with the translation 钢琴 at the formal (lexical), semantic (definitive), and image representation levels.

Table 6: Contrasting *piano*—钢琴

	Lexical level	Definitive level		Image representation	
		noun	verb	object	association
piano (pianoforte)	gentle (soft-loudly)	large instrument, black and white keys, keyboard, small hammers,	演奏 (play) 按 (press)	musical instrument	light (play) gentle (piano)

		metal strings, sound	发出 (produce) 击打 (hit)		
钢琴 (steel-instrument)	钢质乐器 (steel music instrument)	键盘乐器 (keyboard instrument) 钢丝弦 (steel strings) 绒毡 (felt cushion) 木槌 (wooden hammers) 声音 (sound)	按键 (press key) 带动 (trigger) 敲打 (hit) 发出 (produce)	乐器	金属性 (metallic nature) 力度性 (strength)

Table 6 shows that the definitions of *piano* and 钢琴 are quite similar, both describing the physical structure and working mechanism of this musical instrument. The one difference is that in the definition of *piano* the word *play* is used. However, at the lexical level, the two terms differ to a great extent. A comparison of the two terms against the operational definition of lexicographic equivalence in Section 4.2 reveals that the degree of equivalence is low at the lexical level and the world knowledge level. The term *piano* is associated with the timbre of performance (gentle, or soft-loudly when it comes to *pianoforte*). In contrast, the term 钢琴 reveals the material 钢 (steel) of the instrument, thereby generating the “metallic” association in the dictionary user’s mind. Therefore, the terms *piano* and 钢琴 encourage different associations in the

dictionary user's mental representation. In view of this, we have to accept the fact that 钢琴 is not a faithful translation of *piano*, but rather is introduced in the form of a new entity to label the musical instrument.

What, then, in the light of the Whorf hypothesis, is the subtle influence of the instrument's name 钢琴 on the mentality of the individual, and even on the thinking pattern of the entire nation? The name 钢琴 (literally "steel instrument") foregrounds the "steel" composition of the instrument; the 琴键 (key) is "solid" and needs 敲击 (knocking); to play the instrument is to 弹奏 (knock-play); the sound coming from the 钢琴 is associated with "metal" rather than strings. In this way, the "metallic association" (verging on "noise") of 钢琴 is gradually accepted and consolidated, finally "fossilizing" in the mental representation of the language user. As a result, a novice playing the instrument is often found knocking at the keys in a flurry with main force, until the wrists become sore and he totally loses interest in the matter. The amateur audience at a 钢琴 concert might expect, as the name suggests, a "roaring" effect, which they regard as a typical feature of "knocking the steel-instrument."

It is beyond our present discussion as to when the *piano* was introduced to China and how it got the name 钢琴. What we are concerned about is that in the process of introducing terms from or into a foreign language, bilingual dictionary makers need to consider multiple factors, so ensure equivalence from form to content.

Conclusion

The paper set out to explore how insights from philosophical themes and psycholinguistic models contribute to the clarification of lexicographic equivalence. After providing an operational definition of equivalence, two pairs of terms are selected for analysis within the theoretical framework. Through the analysis of the two ways of dictionary translation, that is, looking for equivalent entities (in the case of *terrorism*—恐怖主义) and creating entities (in the case of *piano*—钢琴), the following conclusions can be drawn:

First, both abstract and concrete lexical pairs have non-shared features in the mental representations in their respective cultures, which is in accordance with Quine's indeterminacy of translation, and even with his holism;

Second, bilingual dictionary translation influences the thinking patterns of dictionary users, which is in accordance with Whorf's linguistic relativity hypothesis;

Third, in accordance with the delimitation of equivalence in this paper, lexicographic equivalence is a dynamic process. Total equivalence of dictionary translation can be approached, if not attained.

From the monistic perspective, linguistic form and content are inseparable; they constitute an organized whole; form and content are one, and any change in form brings about change in content. In the final analysis, translational equivalence does not exist. (吴显友 2002) Though the monist's equating form and content is too extreme, it has a grain of truth, in that encoding the same content in different expressions will eventually bring about some change in the content. This echoes Quine's indetermination of translation. For bilingual dictionary makers, this view is undoubtedly a great challenge.

The dictionary defines the standard of language use, and provides the model to follow. What the dictionary compiler puts down in the pages prescribes or guides the dictionary user's understanding and use of words. According to Whorf's linguistic determinism hypothesis, "learning a language changes the way a person thinks" (Carroll 2008: 396). Bilingual dictionary translation influences the thinking pattern of the dictionary user, or the representation of the target word in the user's mental lexicon. The bilingual dictionary maker (or translator) is therefore expected to reduce differences between the two languages, and provide the closest words (or translation), so that the bilingual dictionary user's mental representation of the translated word can approach that of the native speaker.

References

Carroll, David W. 2008. *Psychology of Language*. USA: University of Wisconsin—Superior.

de Groot, A. M. B. 1992. "Bilingual lexical representation: A closer look at conceptual representations". in R. Frost & L. Katz. *Orthography, Phonology, Morphology, and Meaning*: 389 – 412. Amsterdam: Elsevier Science Publishers.

de Groot, A. M. B., Hoeks, John C. J. 1995. "The development of bilingual memory: evidence from word translation by trilinguals". *Journal of Language Learning* 45 (4): 683 – 724.

Hartman, R. R. K. and G. James. 1998. *Dictionary of Lexicography*. London: Routledge.

<http://dictionary.reference.com/browse/terrorism>

<http://en.wikipedia.org/wiki/Terrorism>

Kroll, J. F., & Tokowicz, N. 2005. "Models of bilingual representation and processing: Looking back and to the future". J. F. Kroll & A. M. B. De Groot. *Handbook of Bilingualism: Psycholinguistic Approaches*: 531–553. New York: Oxford University Press.

- Penn, J. M. 1972. *Linguistic relativity versus innate ideas*. The Hague: Mouton.
- Quine, W. 1960. *Word and Object*. Cambridge: the MIT Press.
- Quine, W. 1968. "Ontological relativity". *Journal of Philosophy* 65 (7): 185 – 212.
- Searle, J. 1987. "Indeterminacy, empiricism, and the first person". *Journal of Philosophy*. 86 (3): 123 – 146.
- Szerszunowicz, J. 2015. "Lacunarity, lexicography and beyond: integration of the introduction of a linguo-cultural concept and the development of L2 learners' dictionary skills", *Lexicography ASIALEX* 2:101–118; DOI 10.1007/s40607-015-0015-6
- Van Hell, J. G. & De Groot, A. M. B. 1998. "Conceptual representation in bilingual memory: effects of concreteness and cognate status in word association". *Bilingualism: Language and Cognition*, 1: 193 – 211.
- Wittgenstein, Ludwig. 2009. *Philosophical investigations*. Blackwell Publishing Ltd.
- Zgusta, L. 1971. *Manual of Lexicography*. Publishing House of the Czechoslovak Academy of Sciences.
- Zgusta, L. 2006. *Lexicography Then and Now: Selected Essays*. Edited by Frederic S. F. Dolezal & Thomas B. I. Creamer. Tübingen: Max Niemeyer.
- 陆谷孙. 2012. "飞跃和抵达".《翻译与教学》第二辑, 上海: 复旦大学外国语言文学学院.
- 吴显友. 2002. "文体学中的几个基本问题".《重庆师院学报》(2): 99 – 103.
- Hornby, A. S. 2018. *Oxford Advanced Learner's English-Chinese Dictionary* (9th Edition). Hong Kong: Oxford University Press.
- 汉语大词典编纂处. 2010.《汉语大词典》. 上海: 汉语大词典出版社.
- 中国社会科学院语言研究所词典编辑室. 2016.《现代汉语词典》(第七版). 北京: 商务印书馆.

A PERSPECTIVE ON THE PAST, PRESENT AND FUTURE OF LEXICOGRAPHY WITH SPECIFIC REFERENCE TO AFRICA

D.J. Prinsloo

University of Pretoria

Abstract

The aim of this paper is to give a perspective on the status quo and future of global lexicography and how African lexicography fits into this perspective. The future of dictionaries and in particular the potential threat of internet data to the future of dictionaries will be outlined. Specific aspects of and challenges to African lexicography will briefly be discussed. These aspects include (a) a Euro-centric versus an Afro-centric approach to dictionary compilation, (b) African languages as lesser resourced languages with complicated grammatical systems, (c) confusion in respect of lexicographic traditions and (d) lack of sufficient dictionaries and the absence of a strong dictionary culture. Brief reference will also be made to lexicographic theory. Two basic assumptions which stood the test of time will be taken as basic requirements for African language dictionaries, i.e. sound lemmatisation strategies and sufficient treatment of lemmas — users should be able to find the words and the information about the words that they are looking for in the dictionary.

Key Words: African lexicography, Euro-centric dictionary compilation, Afro-centric dictionary compilation, user support systems, lexicographic traditions

Introduction

Lexicography in Africa does not develop in isolation, it is influenced by the same trends and changes occurring in international lexicography and dictionary user's needs are no different from user's needs in the rest of the world. What is true however is that African lexicography face additional challenges to dictionary compilation in comparison to major languages of the world such as English, German, French, Chinese et cetera. The aim of this paper is to give a perspective on the status quo and future of global lexicography and how African lexicography fits into this perspective. Specific aspects of and challenges to African lexicography in the past and present will be mentioned and some expectations will be outlined for dictionaries of the future and the future of dictionaries for these languages. Three periods of lexicographic development are arbitrarily distinguished. The past period is defined from the compilation of dictionaries on clay tablets hundreds of years ago up to the early nineties at the dawning of the computer era which will

be regarded as the “today” of lexicography and the “tomorrow” as lexicography, say, beyond 2020. In the process the position of lexicographic theory will be considered within the overwhelming emphasis on practical lexicography. User needs and what they require from dictionaries are complicated issues and cannot be justified within the limitations of a single presentation, therefore two basic assumptions which stood the test of time will be taken as a basis. It can be said that users want to find the words and the information about the words that they are looking for in the dictionary.

A good dictionary is one in which you can find the information you are looking for — preferably in the very first place you look. (Haas 1962: 48)

The information to be provided is summarized by Laufer (1992) simply as to guide the user “to know” a word.

Knowing a word would ideally imply familiarity with all its properties. When a person “knows” a word, he/she knows the following: the word's pronunciation, its spelling, its morphological components, if any, the words that are morpho-logically related to it, the word's syntactic behaviour in a sentence, the full range of the word's meaning, the appropriate situations for using the word, its collocational restrictions, its distribution and the relation between the word and other words within a lexical set. Laufer (1992: 71)

African languages as lesser resourced languages

The overview given by the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL, 2014) pinpoints the core characteristics of lesser-resourced languages as follows:

Under-resourced languages suffer from a chronic lack of available resources (human-, financial-, time- and data-wise), and of the fragmentation of efforts in resource development. This often leads to small resources only usable for limited purposes or developed in isolation without much connection with other resources and initiatives. The benefits of reusability, accessibility and data sustainability are, more often than not, out of the reach of such languages.

The number of African languages are estimated to be between 1 250 to 2 100. (Wikipedia, 2017) and many of these languages do not even have a single dictionary. Such languages are often those spoken by poor communities. For many languages there is no standardized orthography.

Grammatical complexity of African languages

African languages are characterised by very complex nominal and verbal systems. The interested reader is referred to Lombard et al. (1985), Poulos and Louwrens (1994) and Van Wyk (1995). Special efforts have to be made to guide users in cases of complex issues where guidance beyond the standard treatment, cross-

references to outer texts or even grammar books are insufficient and specific user support tools are required – see discussion below.

Confusion in respect of lexicographic traditions

A complex interplay exists between lexicographic traditions, lemmatization approaches and lemmatisation strategies. These issues are discussed in detail by Van Wyk (1995) and divided into five categories A to E in table 1 and discussed in great detail in Prinsloo (2009) and Gouws and Prinsloo (2005)

Table 1: Lemmatisation approaches, lexicographic traditions and lemmatisation strategies in Bantu languages

	A Lemmatisation approaches	B Orthography of the language	C Lexicographic traditions	D Lemmatisation strategies : verbs	E Lemmatisation strategies : nouns
1	Traditional	Disjunctive	Stem tradition	Strict stem	Strict stem
2	Rule-orientated	Conjunctive	Word tradition	Left-expanded stem	Left-expanded stem
3	Paradigm				Singular only
4	Frequency				Singular and plural
5					First and 3rd letter

Table 1: Lemmatisation considerations in African Languages (Prinsloo 2009:153)

Low lexicographic quality

Prinsloo (2017:8) is of the opinion that dictionaries for African languages tend to give only very basic treatment of lemmas and that they at best only provide for the most basic text reception needs of the users and are of little use for text production purposes.

Lack of dictionary culture

Atkins (1998a: 3), after having studied the South African situation remarks as follows:

The speakers of African languages have not in their formative years had access to dictionaries of the richness and complexity of those currently available for European languages. They have

not had the chance to internalize the structure and objectives of a good dictionary, monolingual, bilingual or trilingual.

Euro-centric dictionary compilation

Lexicography in Africa is deeply rooted in a Euro-centric approach to dictionary compilation. Euro-centric in this context means that many dictionaries for African languages were compiled by missionaries from Europe to facilitate their work of spreading the gospel. In current literature there is strong condemnation of Euro-centric dictionary compilation – it is equated with colonialism, exploration and even political oppression.

Afro-centric dictionary compilation

What is emphasized and encouraged today is the urge to compile dictionaries for African languages in Africa, by Africans, for Africans (Prinsloo, 2017). The South African government, for example, gave wings to this ideology by establishing state funded National Lexicography Units for all of the 11 official languages of South Africa. Monolingual paper dictionaries are prioritised but bilingual dictionaries and electronic dictionaries are also compiled.

Gouws (2007: 315) links postcolonialism and Afro-centric dictionary compilation as follows:

A characteristic feature of the linguistic situation in the postcolonialization Africa is the reality of emerging indigenous languages. This has led to an increasing need for dictionaries in which these emerging languages are treated. ... This situation created the opportunity for a drastic swing from externally motivated to internally motivated dictionaries, resulting in a situation which sees the majority of new lexicographic projects in Africa characterized by an Afro-centered approach that deviates from the Euro-centered approach.

The current situation is described in more detail in sources such as Gouws (2007), Van Wyk (1995), Prinsloo (2017), Gouws and Prinsloo (2005) and Prinsloo and Taljard (2017).

The challenge to Afro-centric dictionary compilation will thus be to create such sources, and as argued for electronic dictionaries below, especially corpora and lexicographic databases.

Starting afresh with the compilation of e-dictionaries?

The African language lexicographer could assume that good paper dictionaries and a long lexicographic tradition could be regarded as a prerequisite for embarking on the compilation of electronic dictionaries for these languages. Rundell (2012:74), however, regards resources originally developed for printed dictionaries as ‘legacy’ data and says “in an ideal world, we would pulp most of this and start from scratch, producing new resources optimally adapted to digital media”.

This remark tempted African lexicographers to believe that the compilation of good e-dictionaries for African languages is not dependant on the quality of existing paper dictionaries and therefore that they need not start on the back foot by increasing the lexicographic quality of paper dictionaries first, before being able to focus on the compilation of e-dictionaries. The question is whether the compilation of a good e-dictionary can stand in total isolation. Be it as it may, this remark by Rundell brought new hope and motivation that a clean start could be made for the compilation of e-dictionaries for African languages. However, certain major considerations should be honoured right from the beginning, especially that new electronic dictionaries should maximally utilise true electronic features as enabled by the computer era, be corpus based, have appealing screen presentations, pop-up windows and linked to support systems for guidance in respect of especially the salient problematic features of African languages.

Existing e-dictionaries for African languages vary from mere paper dictionaries put “on computer” to wordlists with translations added on, to dictionaries with impoverished treatment of lemmas, to a few dictionaries of good lexicographic quality. See Prinsloo et al. (20:18) for a detailed discussion.

Internet data: friend or foe?

The internet gives access to huge data resources and available information increases exponentially. Dictionary users can get a wealth of information by simply searching for a word or phrase with search engines such as Google, Firefox and Bing. These resources provide exciting opportunities for lexicography, e.g. in the sense of linking internet data to electronic dictionaries but also pose a potential threat to the very existence of dictionaries. Users could find it more useful to simply search the internet for the meaning of a word rather than to consult a paper or electronic dictionary. This threat is real, lexicographers should not merely assume “that it will go away” or that “there will always be dictionaries”. The central theme of Australex 2019 underlines this urgency.

Dictionaries are believed to be under threat, with the shift to digital format and the ongoing undermining of expertise by crowdsourced sites contributing to a perception that dictionaries are no longer needed. Dictionaries are undoubtedly changing and evolving in form and function, but they also remain critically important in providing reliable lexical information and for documenting language. <http://slll.cass.anu.edu.au/centres/andc/australex-2019>

Zimmer (2017) formulates it as follows: “the digital dictionaries ... are also we could say under threat and some of that threat comes from none other than Google”.

Enriching traditional dictionaries with data-driven and visually stimulating features is the most promising way for lexicographers to engage with generations coming of age in the electronic era. If we are truly concerned with allowing learners to appreciate a language’s lexical intricacies, we must supply tools that meet them on their own terrain. Increasingly, this is a

technologized terrain, with expectations of interactivity and fluidity in sophisticated graphic interfaces. (Zimmer: 2017)

The future of dictionaries

In order for dictionaries to survive it is important that they should remain the preferred point of departure for information retrieval. As Zimmer rightfully argues, this objective could be achieved by user-friendly dictionaries enriched through a variety of strategies such as visually stimulating features, and a variety of lexicographic support tools such as linking dictionaries to basic and advanced processed corpus data, writing assistants, tools rendering step-by-step-guidance and verification systems. Consider the following examples where an e-dictionary entry such as *sepela* ‘walk’ is linked to concordance lines, statistical collocation information, spreading of the lemma across sources, and word clusters containing the lemma, extracted from WordSmith Tools (<https://www.lexically.net/wordsmith/>).

sepela (*lediri*) **LEHLALOŠETŠAGOTEE** tsamaya
 1 go iša maoto pele ka go a latelanya mo a dirago gore mmele o tloge mo o bego o le gona; ye ke tiragalo ye bohlokwa mo bophelong ka ge sephedi se kgona go ya mo se ratago : **Bana ba sepela le mo mebileng ba rekiša dilo tša mehuta ye ka moka**
 2 go tloga lefelong le itšeng go ya go le lengwe, e ka ba ka maoto goba ka go šomiša mokgwa wo mongwe wa senamelwa go swana le sefatanaga, bjaloobjalo : **Bošegong bjoo a sepela ka go itlheka a gopotše ga rangwane wa gagwe**
tsamaya (*lediri*) **BONA** sepela

Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete.
 (<https://africanlanguages.com/psl/?timestamp=1557630240000>)

Concordance lines generated for *sepela* as in figure 1:

ka toropong go swan a le ge	0	sepela	le ngwana wa go kgahlwa ke
(mooko 0 mosweu 0 bose)	0	sepela	bjang ka lesobeng le lengwe le
ba ithwesa ka khuru. Motho yola	0	sepela	ka mphaka wa bogale. Re ila
wa gago. Ge a re monna wa gago	0	sepela	le mosadi wa mang- mang, 0
: (0 bolelela fase.) Hleng le yena	0	sepela	gannyane? Tlaa re mo sie! (0 a
e nyakile go ba gata. Bjale mmotoro	0	sepela	gabotse gape. Ba bona
go laola digoba tsa gagwe gomme	0	sepela	ka go tshetsherega. Alekoholo
toropong mediro ke ye mentsi. Piti	0	sepela	le tatagwe ge a lok isa ditaba
ba ithwesa ka letolo. Motho yola	0	sepela	ka thipa ya bogale. Re ila ra
haba ka boithatelo le boroto Ngwedi	0	sepela	ka go kgokgona le bodutu,

Figure 1: An extract of the concordance lines generated for *sepela*

Words co-occurring with *sepela*, their frequencies and positions to the left and right of *sepela*, figure 2:

Word	With Relation Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1
SEPELA	sepela 0.000	641	5 723	110	110	11	31	20	42	6	5 503	6
GABOTSE	sepela 0.000	115	210	27	183	7	10	3	5	2		168
GONA	sepela 0.000	118	201	87	114	24	21	19	17	6		26
GOMME	sepela 0.000	107	179	91	88	17	16	17	41			30
YENA	sepela 0.000	75	151	66	85	15	15	17	13	6		2
TŠA	sepela 0.000	67	140	83	57	18	17	9	15	24		1
FELA	sepela 0.000	91	138	63	75	11	14	9	27	2		17
WENA	sepela 0.000	64	131	60	71	9	7	9	12	23		19
BJALO	sepela 0.000	89	129	46	83	13	15	6	8	4		46

Figure 2: An extract of collocations of *sepela*

Graphical illustration of the spreading of *sepela* across a number of sources, cf. figure 3:.



Figure 3: A section of the graphic presentation of the spreading of *sepela* across sources

Three-word clusters with *sepela*, cf. figure 4.

Cluster	Freq.	Set	Length
A SEPELA A	234		3
GO SEPELA KA	192		3
A SEPELA KA	171		3
GO SEPELA LE	148		3
A SEPELA LE	140		3
GE A SEPELA	139		3
SEPELA KA GO	127		3
BA SEPELA BA	120		3

Figure 4: A section of three-word clusters with *sepela*.

Activities to produce Words Sketches (<https://www.sketchengine.eu/user-guide/user-manual/word-sketch/>) for African languages are ongoing but dependant on large tagged corpora.

A sophisticated writing assistant, the Sepedi Helper in figure 5 (<http://www.sepedihelper.co.za>) has been compiled to assist users with the verbal moods. The user is prompted to insert basic nouns and verbs and the writing assistant construct phrases and sentences.

Sepedi:
 Prototype sentence builder [? Instructions](#)

Sentence Type: **Indicative** [?](#)

STATEMENTS

Sentence:
 [Choose Subject Noun] [Choose Verb]

Tense: **Present** [?](#) Type: **Positive** [?](#)

Input List:

- Tense** - Present
- Type** - Positive
- Subject Noun** - [Not Chosen]
- Verb** - [Not Chosen]
- Direct Object Noun** - [Not Chosen]
- Indirect Object Noun** - [Not Chosen]

Created by: DJ Prinsloo, Daniel Prinsloo and Jacobus Prinsloo
 Developed for SeLA: Scientific e-Lexicography for Africa

[? About](#)

Figure 5: A writing assistant for the verbal moods in Sepedi

A step by step selection tree is under development for copulative constructions in Sepedi. This decision tree guides the user through the complicated copulative structure in this language to find the correct Sepedi copulative. See figure 6

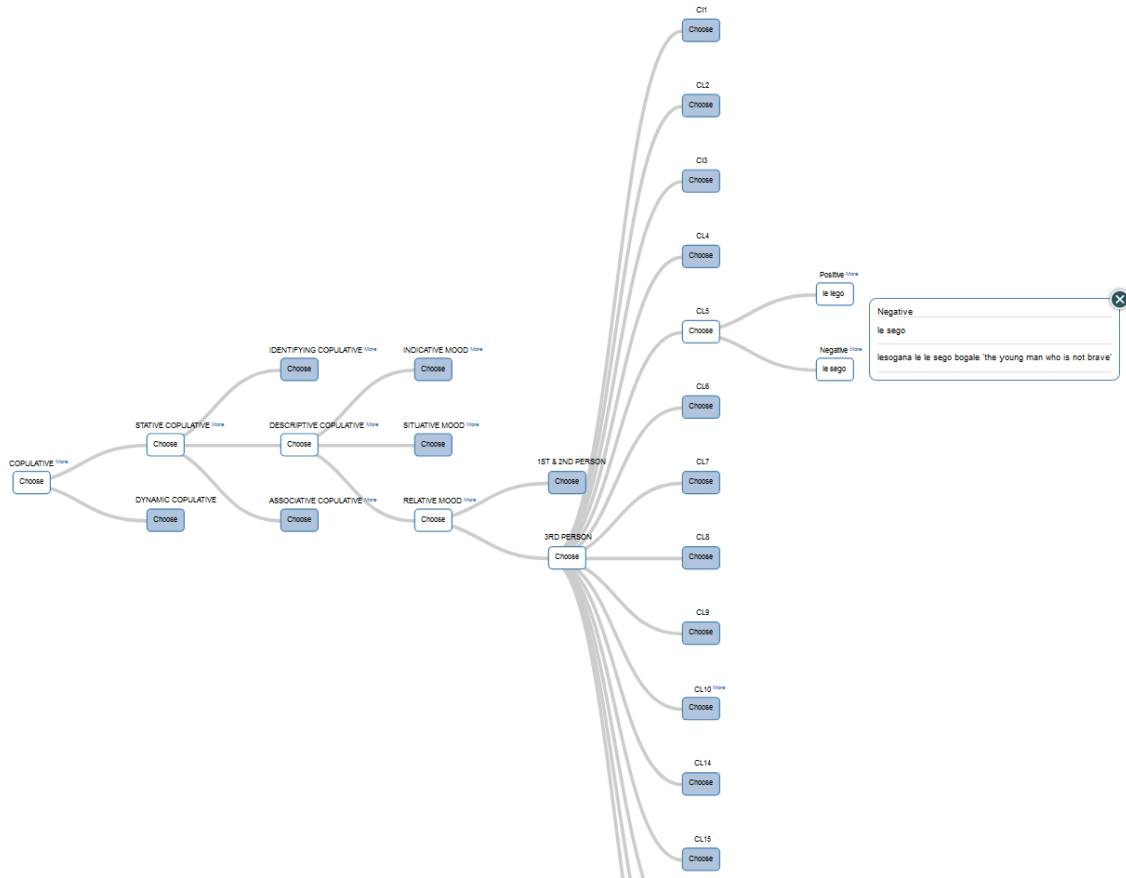


Figure 6: The Sepedi copulative selection tree

Finally, consider the recently developed *Write Assistant*. According to its self-description *Write Assistant* is an integrated spelling aid, thesaurus and translation tool activated from within a word processor. (<http://www.writeassistant.com/en/>)

The conference theme of eLex 2019 is quite appropriately formulated as “**smart lexicography**” with special reference to dictionaries on mobile devices and to maximizing use of dictionary content and linking of dictionaries to support tools.

In the last decade, the use of mobile devices such as smartphones has increased considerably. ... smart use and reuse of dictionary content. ... have them linked to other dictionaries and language resources, or even integrated in various tools.

(<https://elex.link/elex2019/>)

Lexicographers should seriously consider the so-called 15 Villa Vigoni-Theses on Lexicography on the EMLex website (<https://www.emlex.phil.fau.eu/other/>). Under the heading *Dictionaries for the Future* –

The Future of Dictionaries. Challenges to Lexicography in a Digital Society specific guidelines were formulated such as:

- Dictionaries of the future are lexical or linguistic information systems in which existing lexicographic data are conflated.
- Practical lexicography must constantly be aware of its social responsibility and must strive for a comprehensive, pluralistic description of linguistic and factual realities.
- One significant task for the lexicography of the digital future is the orderly conflation of data which has been generated automatically by text corpora and specifically processed as well as a user-orientated presentation.
- Lexicographic projects should be oriented towards the specific needs of the users.

The status and role of lexicographic theory

Gouws and Prinsloo (2005:1) say that “the theoretical component can be regarded as a relative late-comer because lexicography has originally only been associated with the practice of dictionary-making”. According to them the form, contents and functions of dictionaries constitute the focus of the theoretical component and the practical component is the compilation of dictionaries. Rundell (2012) refers to the existence of an uneasy relationship between practical and theoretical lexicography and states that lexicography works in practice but asked whether it also works in theory. There are differences of opinion regarding lexicographic theory and its role in relation to practical lexicography even up to the question whether lexicographic theory exists at all. Bergenholtz and Gouws summarize the situation as follows.

... Atkins and Rundell (2008: 4) saying, with regard to a theory of lexicography, that they "do not believe that such a thing exists", and Bejoint (2010: 381) saying: "I simply do not believe that there exists a theory of lexicography, and I very much doubt that there can be one", to lexicographers who firmly believe in a lexicographic theory, cf. Wiegand (1989), Bergenholtz and Tarp (2003), Gouws (2011), Tarp (2012). Bergenholtz and Gouws (2012:36)

The challenge posed to theoretical lexicography is to lead the way in the sense that lexicographers can compile better dictionaries by following theoretical guidance.

Dictionaries of the future

It is expected that paper dictionaries for African languages will remain the preferred option for many years to come but it will be paralleled by the development of electronic dictionaries using true electronic features enabled by the computer such as pop-up windows, audible pronunciation and a variety of dictionary support tools. It is also expected that the Afro-centric approach will gain momentum rendering more dictionaries of lexicographic achievement and user-generated content through community engagement. It is believed that the internet will play a vital role in future information retrieval. Rundell (2012:83) says:

The current situation is messy, with a great deal of interesting but uncoordinated activity, and plenty of trial and error. ... the situation continues to change rapidly, as technologies from the wider field of Internet search increasingly impact on what we do. ... But the model which has served us so far still looks serviceable: the basic principles of focussing on the user and being faithful to the language data; seeking guidance from relevant linguistic and computational theory; and drawing on good-quality user research to identify what works.

Conclusion

In this presentation it was attempted to give a perspective on the status quo and future of global lexicography and how African lexicography fits into this perspective. It was indicated that African languages, although subjected to the same challenges as other languages of the world, have their own additional challenges. These challenges mainly revolve around socio-economic and political problems, data scarceness, complicated grammatical structures, lack of a dictionary culture, etc. as outlined above. Viewed from a positive angle there are driving forces such as an Afro-centric approach, community engagement, governmental initiatives, Human Language Technology projects, national lexicography units, support from publishers, lexicographic research, and lexicographic associations of the “-lex family” , such as Afrilex, Euralex, Australex, Asialex and Globalex.

Acknowledgements

This research is supported in part by (a) the South African Centre for Digital Language Resources (SADiLaR) and (b) the National Research Foundation of South Africa (Grant specific unique reference numbers 85763) The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are those of the authors, and that the sponsors accept no liability whatsoever in this regard.

References

- Atkins, B.T.S. 1998. Some discussion points arising from Afrilex-Salex'98. *Unpublished course evaluation document*, University of Pretoria.
- Australex 2019. Lexicography and dictionaries in the public sphere. Available online at <http://slll.cass.anu.edu.au/centres/andc/australex-2019>.
- Bergenholtz, H. and Gouws, R.H. 2012. What is Lexicography? *Lexikos* 22: (2012): 31-42
- CCURL. 2014. Proceedings overview: *Workshop on collaboration and computing for under-resourced languages in the Linked Open Data Era*. Available online at <http://www.ilc.cnr.it/ccurl2014/>.
- eLex 2019 – electronic lexicography in the 21st century. Available online at <https://elex.link/elex2019/>.
- EMLex, *Villa Vigoni-Theses: Dictionaries for the Future – The Future of Dictionaries. Challenges to Lexicography in a Digital Society*. Available online at <https://www.emlex.phil.fau.eu/other/>.
- Gouws, R.H. 2007. On the development of bilingual dictionaries in South Africa: aspects of dictionary culture and government policy. *International Journal of Lexicography*, 20(3): 313–327.
- Gouws, R.H. and Prinsloo, D.J. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: African Sun Media.
- Haas, M.R. 1962. *What Belongs in the Bilingual Dictionary?* In: Householder, F.W. and S. Saporta (eds.) 1962: 45-50.
- Laufer, B. 1992. Corpus-based versus Lexicographer Examples in Comprehension and Production of New Words. In: Tommola, H. et al (eds.) 1992. *Euralex '92 Proceedings*. Tampere: University of Tampere: 71-76.
- Lombard, D.P., Van Wyk, E.B. and Mokgokong, P.C. 1985. *Introduction to the Grammar of Northern Sotho*. Pretoria: Van Schaik.
- Poulos, G. and Louwrens, L.J. 1994. *A Linguistic Analysis of Northern Sotho*. Pretoria: Via Afrika.
- Prinsloo, D.J. 2009. Current Lexicography Practice in Bantu with Specific Reference to the Oxford Northern Sotho School dictionary. *International Journal of Lexicography*. 2009; Oxford University Press (UK) 22(2): 151-178.
- Prinsloo, D.J. 2017. Analyzing words as a social enterprise: Lexicography in Africa with specific reference to South Africa. In J. Miller (Ed.), *Analysing words as a social enterprise: Celebrating 40 years of the 1975 Helsinki Declaration on lexicography*. Available online at <https://www.adelaide.edu.au/australex/publications/>.
- Prinsloo, D.J., Prinsloo, J.V. and Prinsloo, Daniel. 2018. African Lexicography in the Era of the Internet. *The Routledge Handbook of Lexicography*. Pedro A. Fuertes Olivera (Ed.). London: Routledge. 487-502.

Prinsloo, D.J and Taljard, E. 2017. Afrikataalleksikografie: gister, vandag en môre [African Language Lexicography: yesterday, today and tomorrow]. *Lexikos* 27. (427-456).

Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete. Available online at <https://africanlanguages.com/psl/>?

Rundell, M. 2012. It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In Ruth Vatvedt Fjeld en Julie Matilde Torjusen (Eds.). *Proceedings of the 15th Euralex International Congress*. 7-11 August 2012. Oslo.

Sepedi Helper. Available online at <http://www.sepedihelper.co.za>

Van Wyk, E.B. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. In: *Lexikos* 5: 82-96.

Wikipedia. 2017. Languages of Africa. Available online at https://en.wikipedia.org/wiki/Languages_of_Africa#/media/File:Map_of_African_language_families.svg.

WordSmith Tools Available online at <https://www.lexically.net/wordsmith/>.

Zimmer, B. 2017 Defining the Digital Dictionary: How to Build More Useful Online Lexical References. Keynote, *Fifth biennial conference on electronic lexicography, eLex 2017*, Leiden, Netherlands, 19-21 September 2017.

CONSIDERATIONS FOR PROVIDING ETYMOLOGICAL INFORMATION IN THE KBBI INDONESIAN DICTIONARY

David Moeljadi

Palacký University Olomouc

Ian Kamajaya

ASTrio Pte Ltd

Azhari Dasman Darnis

Badan Bahasa

Abstract

We discuss the inclusion of etymological information in the Indonesian dictionary KBBI (Kamus Besar Bahasa Indonesia) fifth edition. KBBI is the most comprehensive and authoritative Indonesian monolingual dictionary, published/launched by The Language Development and Cultivation Agency, under the Ministry of Education and Culture (Moeljadi et al. 2017, Kamajaya et al. 2017). It is mainly online-based (<https://kbbi.kemdikbud.go.id>), updated regularly, and will be enriched by etymological information from October 2019. This etymological information is valuable for the Indonesian language that has loanwords from various languages and language families: Austronesian (Old Javanese), Indo-European (Sanskrit, Persian, Portuguese, Dutch, English), Dravidian (Tamil), Semitic (Arabic), and Sinitic (Hokkien, Cantonese, Mandarin). The first etymology project began in the late 2000s, after the fourth edition was published. A team was formed based on the grouping of donor languages: (1) Arabic and Semitic languages, including Persian; (2) European languages (Dutch, English, and French); (3) Old Javanese; and (4) Chinese

languages. Unfortunately, this project did not meet the target. Starting from the second etymology project, since 2016 each year we focus on two groups and involve experts from universities in Indonesia. We refer to previous work and references e.g. Jones et al. (2007). Etymology projects on Sanskrit and Old Javanese were carried out (2016~2017), after that Dutch (2017~) and Arabic (2018~). Data collection is based on the dictionary headwords. The technical part involves programming and database restructuring. The existing KBBI database will be augmented with etymological information-related tables containing the original scripts of the loanwords and the relationships within them as well as between them and the entries, enabling KBBI to present etymological relations of the loanwords accurately. We believe that the etymological information in KBBI will serve as a valuable resource and accompaniment to Indonesian historical, linguistics, and lexicography research.

Key Words: Indonesian, etymology, online dictionary, database structure

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. Its first edition, published in 1988, has 62,000 entries. The number of entries increased to 72,000 or about 10,000 entries over three years in the second edition (1991). Its third edition, published in 2001, contains 78,000 entries and seven years later, the number of entries in the fourth edition increased to more than 92,000. Its latest, fifth edition was released for the first time in 2016 in three formats: printed, online, and offline mobile versions. Since then, it is

regularly updated twice a year. As of April 2019, it has more than 110,000 entries. Moeljadi et al. (2017) describes the creation of the database as well as the database structure. Kamajaya et al. (2017) explains the online KBBI in details. Although KBBI has comprehensive data on headwords, derived words, compounds, proverbs, word senses, parts-of-speech etc., it does not have information on the word origins or etymology: from which language(s) a particular word is borrowed, the etymological route it takes until it enters the Indonesian language, as well as changes in word forms, sounds, and meanings. This paper discusses the inclusion of this etymological information into KBBI database, especially regarding the database structure.

Historical background and contact with other languages

For centuries the ancestors of Indonesian speakers had a contact with speakers from various nations in the world. Sanskrit is recorded as the first foreign language which entered Indonesia, since the beginning of A.D. This language became the language of literature and also a medium of spread of Hinduism and Buddhism. Hinduism spread vastly in Java in the seventh and eighth century, then Buddhism in the eighth and ninth century. Together with the spreading of Hinduism, the spice trade with Indians also took place. Some of the Indians are Hindi speakers, while some are Tamil speakers from Southern India and Eastern Sri Lanka, whose language became the medium of literary work. Tamil language had a strong influence on Malay language.

Contact with Chinese speakers had happened since the seventh century, when Chinese merchants traded to Riau Islands, West Kalimantan, and East Kalimantan, even until North Maluku. 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. What is called

“Chinese” here, actually is more accurately said as languages from China. There are many languages in China. Four of them are well-known languages in Indonesia, i.e. Hokkien, Hakka, Cantonese, and Mandarin. Because the contact had lasted long, it is reasonable that many loanwords from Chinese languages came into Malay/Indonesian. However, because the Chinese languages were not used as religious, scientific, or literary medium in Indonesia, it is very likely that many of those loanwords blended into languages in Indonesia.

The Arabic language was brought to Indonesia from the seventh century by merchants from Persia, India, and Arab who were also the spreaders of Islam. Arabic, an Islamic religious language, began to influence Malay, especially since the twelfth century when many kings embraced Islam. Because many of those merchants were Persian speakers, quite a number of Persian words entered into Malay.

The Portuguese language had begun to be known by Malay-speaking community since the Portuguese occupied Malacca in 1511 after they occupied Goa one year before. Portuguese were unable to compete with the Dutch who came later. They stepped aside to the eastern region of the Archipelago. However, in 17th century Portuguese, together with Malay, had become a lingua franca among ethnic groups in the Archipelago.

The Dutch people started to come to the Archipelago in the beginning of 17th century when they expelled Portugues from the Moluccas (Maluku) in 1606 and then headed to the Java island and other regions in the west. Since then, gradually the Dutch people took control of many regions in Indonesia. The Dutch language could not completely oust the Portuguese language because Dutch was more difficult to learn and Dutch people did not like to open up to people who wanted to learn Dutch culture, including the language. However, the Dutch occupation was gradually covering almost the entire country and lasted for a long time. The Netherlands was also a main source of

learning for youth groups in the independence movement. Because of that, the concept of state building mostly refers to the Dutch language.

The British once occupied Indonesia although not for long. Raffles invaded Batavia (now Jakarta, the capital city of Indonesia) in 1811 and stayed there for five years. Before he was moved to Singapore, he also stayed in Bengkulu in 1818. Actually, in 1696 the British once sent a messenger Ralph Orp to Padang but landed in Bengkulu and settled there. In Bengkulu the British built a fort named Fort Marlborough in 1714-1719. It means that more or less a contact with English had already happened for a long time in a region near the center of Malay-speaking community. However, the most intensive contact with English happens in this globalization era.

Several years before the independence in 1945, Japan occupied Indonesia. However, the Japanese occupation lasted only for three and a half years and it left very few words which can survive through generations.

Because of this historical background and the richness of loanwords in Indonesian, many etymological work have been done by various researchers, such as Jones (2007) and Tadmor (2009). In Indonesia, some small-scale research has been done before the year 2000, such as *Kamus Etimologi* “Etymological Dictionary” (Harun et al. 1984) that was originally a research report compiled by a research team, it contains only words borrowed from Arabic; *Kamus Etimologi Bahasa Indonesia* “Indonesian Etymological Dictionary” (Adiwimarta et al. 1987), one of the results of Jakarta project research in 1984-1985 which was done by one team of researchers; and *Senarai Kata Serapan dalam Bahasa Indonesia* “A List of Loanwords in Indonesian” (Jumariam et al. 1996). However, to the best of our knowledge, there is no comprehensive work for Indonesian loanwords that has an open-source database that can be used to augment the existing KBBI database.

Method and Results

The idea of the inclusion of etymological information in KBBI was proposed to the KBBI's editorial board in the early 2000s. This is in line with one of the recommendations from the Congress of the Indonesian Language in 2003, i.e. to provide etymological information in KBBI, since the information of the word origins in KBBI is very simple and limited. KBBI provides only information about the donor language for some headwords, for example the word “kamsia” is labeled *Cn* (Chinese), without its original form, sound, and meaning in Chinese.¹²

In 2010 the first etymological project of KBBI started. The main purpose is to provide KBBI with reliable etymological information, i.e. the origin of the words, the changes both in the sense and form, and the etymological route each word took before it was borrowed into Indonesian, e.g.

butik *n* toko tempat menjual pakaian jadi dengan segala kelengkapannya (terutama untuk wanita)

[< Bld *boutiek* 'toko kecil, eksklusif, menjual barang/mode mewah; rumah mode' < Pr *boutique* 'bagian depan rumah atau gedung tempat penjual, perajin memamerkan dan menjajakan, serta menjual dagangannya; warung, kedai, toko' < Yn *apothēkē* 'gudang penyimpanan sediaan']

In the beginning, the project involved lexicographers from Pusat Bahasa.¹³ None of them had the expertise in etymology. They relied on dictionaries and etymological research reports. One of the

12 <https://kbbi.kemdikbud.go.id/entri/kamsia>

13 Pusat Bahasa is the former name of Badan Bahasa (The Language Development and Cultivation Agency).

reports is from the research conducted by Pusat Bahasa under the title *Senarai Kata Serapan dalam Bahasa Indonesia* (SKSBI) “A List of Loanwords in Indonesian” (Jumariam et al. 1996). Another one is from an Indonesian Etymological Project conducted by KITLV, Leiden, summarized in a book “Loan-Words in Indonesian and Malay” (LWIM) (Jones 2007). The latter becomes the standard of how the etymological information will be presented in KBBI.

The project began with the list of entries in SKSBI and LWIM. SKSBI lists words in KBBI borrowed from 11 languages: Arabic, Dutch, Chinese, English, Portuguese, Sanskrit, Old Javanese (OJ), Hindi, Persian, Tamil, and Malay. Entries were listed separately based on the donor languages and compared with the ones in LWIM and with dictionaries of donor languages. The words of English origin, for example, were verified using Merriam-Webster Dictionary, Oxford Dictionary, Chambers Dictionary of Etymology, Dictionary of English Etymology, and Microsoft Encarta 2009. The lemma *carter* in SKSBI (page 96) is from English “charter”, its meaning is “the hiring of something for a special purpose”. In LWIM, it is written “**carter** [charter; rent, let] < Eng *charter*”. Afterwards, it is compared with other sources to check the original form and meaning, as well as the etymological route it took before entering Indonesian.

carter *v* [< Ing *charter* 'menyewa kendaraan untuk tujuan pribadi atau khusus' < Pr *chartre* < Lt *chartula* < *charta* 'kertas, daun papirus']

However, because of some budget and priority issues, in 2012 the project was pushed aside and had to take a break for a year. The result of this two-year work is 398 lemmas with etymological information. Most of them have the information only on original forms and senses, without etymological routes. After a one-year break, the project started again.

In 2013 some changes and improvements were made. Some experts were involved. Experts of Arabic, Dutch, Chinese, and Persian from several universities were asked to verify the work done

in the previous project as well as to complete the unfinished ones. The project run until 2015 because in 2015 a project of building a KBBI database and online KBBI was given more priority. The result of this four-year work is shown in Table 1.

Table 1: Result of KBBI etymological project 2010-2012, 2013-2015

	Arabic	Dutch	English	Chinese	OJ	Korea	Sanskri n	Persia t	Japanes n	Japanes e
Alphabet	a, i, j, k, l, m, n, r, s, t	a, f, h, i, l, m, p, q, r, s	c, d, e, f, g, h, I, j, k, p, r, s	all	a--y	all	all	all	all	all
Not-edited	500	846	824	-	1156	64	76	31	275	
Edited	850	-	-	432	-	-	-	-	-	-
Verified	-	-	-	-	-	-	-	-	-	-

The etymological information for loanwords from Old Javanese, Korean, Sanskrit, Persian, and Japanese is limited. It does not include the etymological routes. Only Chinese and some of Arabic data was edited.

The etymology project that stopped in 2015 resumed a year later in 2016. It focused on: (1) editing and completion of etymological information of loanwords from Arabic, Sanskrit, Old Javanese, and Dutch; (2) verification of etymological information of loanwords from Chinese, Sanskrit, Old Javanese, and Arabic; and (3) preparation to build a database system. To complete these objectives, a number of experts was involved. Language experts focused on the completion, editing, and verification of etymological information. Information technology experts focused on building a database system. In order to build the database, the data has to be written in a specific format. The online KBBI is planned to be updated in October 2019 with etymological information of loanwords from Arabic. In 2020 it will be updated with information on loanwords from Chinese, Dutch, Sanskrit, and Old Javanese. Table 2 shows the result of work from 2016.

Table 2: Result of KBBI etymological project 2016-present (April 2019)

	Arabic	Dutch	English	Chinese	OJ	Korean	Sanskrit	Persian	Japanese
Alphabet	a, i, j, k, l, m, n, r, s, t	a, f, h, i, l, m, p, q, r, s	c, d, e, f, g, h, I, j, k, p, r, s	all	a--y	all	all	all	all
Not-edited	-	-	824	-	-	64	-	31	275
Edited	1450	1938	-	-	303	-	932	-	-

The work on loanwords from Arabic, Dutch, Old Javanese, and Sanskrit has arrived at the editing stage. Loanwords from Chinese have been edited and the verification process is in progress. Loanwords from English, Korean, Persian, and Japanese are going to be edited and verified next year due to scale of priorities and shortage of budget. Some words which are regarded as of Old Javanese origin in Table 1, were edited and regarded as of Sanskrit origin in Table 2. Old Javanese and Sanskrit experts distinguished loanwords from Old Javanese and Sanskrit, especially the ones which have similar forms and senses. The project is still continuing until all loanwords are verified. It is planned to end in 2020. At the final stage, the focus is on the data input of the etymological information into KBBI. The project will result in 5,012 loanwords with complete etymological information to enrich the lemmas in KBBI.

Discussion

This section discusses the KBBI database restructuring/augmentation for etymological information. We build three additional tables as follows:

1. Etymological Source Table, consists of a list of sources/references for etymological information.
2. Etymon Table, consists of a list of etymons together with the information on donor languages, original senses, linguistic information (e.g. part-of-speech, person, number, gender, tense, aspect, voice), and sources/references.
3. Entity Table, represents a ‘minimally-designed’, ‘properties-augmented’ abstract object called

‘entity’ which we build mainly to represent complex logical-etymological relationships between etymons and headwords in KBBI.

Each table is discussed in the next subsections.

Etymological Source Table

This table consists of a list of etymological sources, their identifiers (IDs), and notes. At present, it has three columns as follows:

1. Etymological Source ID, consists of one unique ID for each source
2. Etymological Source, consists of source names (books, dictionaries, reports etc), author names, years, publishers
3. Notes, consists of optional notes or descriptions for each source

Etymon Table

Each row in this table basically consists of a set of an etymon, its donor language, and its original sense. Each set has a unique identifier. It is possible that two or more etymons have exactly the same form but come from different languages of origin or they are homonyms, having different senses. The unique identifier will resolve the ambiguity. This table consists of 14 columns as follows:

1. Etymon ID, consists of one unique ID for each set of etymon, donor language, and sense
2. Etymon (original script), consists of a word or morpheme written in the original script, from which a later word is derived
3. Etymon (transliteration), consists of a word or morpheme transliterated or written in Latin

script if the original script is not in Latin script

4. Donor Language, consists of the language of origin
5. Sense, consists of the original meaning of the etymon
6. Etymological Source ID, consists of a set of etymological source IDs. These IDs are linked to the ones in the etymological source table.
7. Part-of-speech, consists of the part-of-speech (noun, verb, adjective etc.) of the etymon
8. Person, consists of information on first, second, and third person, if relevant
9. Number, consists of information on number, such as singular, dual, and plural, if relevant
10. Gender, consists of information on gender, such as masculine, feminine, and neutral, if relevant
11. Tense, consists of information on tense, such as past, present, and future, if relevant
12. Aspect, consists of information on aspect, such as perfect, imperfective, and progressive, if relevant
13. Voice, consists of information on voice, such as active, passive, and medial, if relevant
14. Case, consists of information on case, such as nominative, genitive, dative, and accusative, if relevant

Entity Table

The last, and perhaps the most complicated, step we do is to build the ‘entity’ table. We need to build a ‘building-block unit’ to represent etymological relationships between one etymon and another etymon, as well as between one etymon and one headword. We use a certain design philosophy, i.e. to build ‘minimum’ building-block units which are capable of encompassing all the ‘logical-etymological’ relationships. We call this minimum building-block unit ‘entity’. Therefore, an ‘entity’ is a ‘minimally-designed’, ‘properties-augmented’, abstract object,

consisting of one reference of etymon and logical-etymological relation to two entities and/or one headword. This table consists of 7 columns, as follows:

1. Entity ID, consists of one unique ID of entity for each row
2. Headword ID, consists of headword ID in the existing KBBI database that is directly related to the respective entity. This ID serves as the primary link to the current KBBI database.
3. First reference entity ID, consists of the reference ID of an entity that is the base of the respective entity.
4. Second reference entity ID, consists of the reference ID of another entity that is the base of the respective entity.
5. Logic, consists of the logical-etymological relationship between the respective entity and its 'referenced entities'. At present, there are five possible values: ROOT, AND, OR, A-AND, and A-OR.
 - a. Logic ROOT makes use of the value only in the first reference entity ID column (the second column is left blank) and is used if the first reference entity is the root for the respective entity.
 - b. Logic AND is used if the first and the second reference entities are combined to form the respective entity.
 - c. Logic OR is used if either of the reference entities is a possible root for the respective entity.
 - d. Logic A-AND, stands for 'Augmented-AND', is used to augment one more AND logic in the same level between two reference entities, one of which must already have AND or A-AND logic. This is essentially used to create three or more same-level AND relationships between entities, which would not be possible to be represented by having only two entity reference columns. Note that although logically (1) (X AND Y) AND Z, (2) X AND (Y

- AND Z), and (3) X AND Y AND Z have the same result, they are not etymologically identical. For case (1), X AND Y form a word P and then P AND Z form a word W. For case (2), Y AND Z form a word Q and then X AND Q form a word W. For case (3), X AND Y AND Z together form a word W. Adding another A-AND in reference to an entity which already has A-AND will result in augmented AND in the same possible lowest level. Suppose an entity W is formed from X A-AND (Y AND Z), i.e. $W = X \text{ AND } Y \text{ AND } Z$, referencing an entity V with the entity W by using A-AND will result in V A-AND (X AND Y AND Z), i.e. $V \text{ AND } X \text{ AND } Y \text{ AND } Z$. It can be easily seen that this A-AND logic allows augmentation of as many AND logic in the same level as possible.
- e. Logic A-OR, stands for ‘Augmented-OR’, is used to augment one more OR logic in the same level between two reference entities, one of which must already have OR or A-OR logic. The use of this logic is the same as its A-AND counterpart, i.e. to create a same-level OR relationship between three or more entities.
6. Etymon ID, consists of the ID of the set of etymon, donor language, and sense in the Etymon table related to this entity
7. Notes, consists of explanations or descriptions on the history of usage etc.

The entity is said to be ‘minimally-designed’ because (1) the number of referenced entities cannot be more than two and (2) the number of properties to be augmented to an entity are kept to be as few as possible. As explained above for A-AND and A-OR, referencing two entities is minimum to create all possible logical-etymological references. We use the term ‘logical-etymological’ because the reference between entities are both logical (having OR or AND) and etymological (as explained in the use cases of A-AND). Although an entity may not have any logical-etymological reference with other entities, it will always have at least one logical-etymological relationship with another entity. An entity may not have any logical-etymological reference with other entities

because it is a base for other entities. Other entities will be referenced to that entity. Table 3 illustrates the relations between entities in Entity Table. In this table, the ‘Headword ID’ is replaced by ‘Headword’ for clarity purpose. ‘Etymon ID’ is replaced by ‘Etymon’, ‘Donor Language’, and ‘Sense’. ‘Etymon (original script)’ and ‘Etymon (transliteration)’ are merged in one column ‘Etymon’ to save space.

Table 3: An example of Entity Table

Entity ID	Headword	Ref. Entity ID 1	Ref. Entity ID 2	Logic	Etymon	Donor Language	Sense
1					تألّه (ta'allaha)	Arabic	To worship
2	ilah	1		ROOT	إله (ilāh)	Arabic	One who is worshipped
3	Allah	2		ROOT	الله (Allāh)	Arabic	One who is worshipped
4					خلف (khalafa)	Arabic	To replace or to represent

5	khalifah	4		ROOT	خليفة (khalīfah)	Arabic	Representative or leader
6	khalifatullah	3	5	AND	خليفة الله (khalīfatullah)	Arabic	God's representative
7					خار-يخير (khāra)	Arabic	Exceeding, the better one
8					خار-يخور (khāra)	Arabic	Mooing (animal voice)
9		7	8	OR	استخار (istakhāra)	Arabic	To request for mercy
10		7		ROOT	اختار (ikhtāra)	Arabic	To choose from
11	istikharah	9		ROOT	استخارة (istikhārah)	Arabic	To request to be given the best among two or more choices
12	akhir				آخر (ākhir)	Arabic	Behind, eternal essence

From the example above, we can make these distinctions:

1. ID 1, 4, 7, 8, and 12 are called ‘basic entities’. These entities do not have any logical-etymological reference with other entities.
2. ID 2, 3, 5, 6, 9, 10, and 11 are called ‘composite entities’. These entities have at least one logical-etymological reference.
3. ID 2, 3, 5, 6, 11, and 12 are called ‘head entities’. These entities are directly related to headwords.

In addition to database structure, there are several things that need to be considered in the verification stage, such as reliable data sources, standardized transliterations, and scripts (non-Unicode characters, variants).

Conclusion

We have discussed the augmentation of KBBI database with etymological information. The aim is to augment the existing KBBI database with complex etymological information, especially etymological routes and logical-etymological relationships. We have designed a database structure with tables that can accommodate every possibility of etymological routes and logical-etymological relationships. However, there are some issues left, such as verification of sources by experts and workflow towards the end of the project (data input).

Acknowledgment

We thank Ibnu Kharish for helping us with loanwords from Arabic. We 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.

References

Adiwimarta, Sri Sukei, Adi Sunaryo, Saodah Nasution, Hartini Supadi, Achmad Patoni, and Umi Basiroh (1987) *Kamus Etimologi Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Harun, Ramli, Aliudin Mahyudin, and Achmad Patoni (1984) *Kamus Etimologi Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Jones, Russell (general ed.), C.D. Grijns, J.W. de Vries (eds.) (2007) *Loan-words in Indonesian and Malay*. Compiled by the Indonesian Etymological Project. Leiden: KITLV Press.

Jumariam, Meity T. Qodratillah, and C. Ruddyanto (eds.) (1996) *Senarai Kata Serapan dalam Bahasa Indonesia*. Jakarta: Departemen Pendidikan dan Kebudayaan.

Kamajaya, Ian, David Moeljadi, and Dora Amalia (2017) KBBI Daring: A Revolution in The Indonesian Lexicography. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.*, Leiden, pp 513–530.

Moeljadi, David, Ian Kamajaya, and Dora Amalia (2017) Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications. In *Proceedings of The 11th International Conference of the Asian Association for Lexicography*, Guangzhou, pp 64–80.

Tadmor, Uri (2009) Loanwords in Indonesian. In Haspelmath, Martin and Tadmor, Uri (eds.) *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

SIGNIFICANCE OF PHRASEOLOGY FOR LEXICOGRAPHY AND PHRASE MINING BASED ON CHINESE CORPUS

Dejun Li

Man Fu

(College of International Studies, National University of Defense Technology)

Abstract

The value of corpus to lexicography is not to be overestimated. Nowadays, corpus plays significant roles not only in selecting citations but also in writing dictionary definitions. However, drawbacks of corpus-driven method should not be neglected. In discussion of corpus-driven methodology in this paper, concrete examples are cited to elucidate its validity. However, because of its heavy reliance on interpretation of concordance lines in Chinese lexicography, current corpus-driven methodology possesses some drawbacks which are hard to avoid if there is no other way out.

Phraseologies in corpus linguistics are co-occurrences of two or more tokens which function as an independent semantic unit and with statistical significance. They include not only fossilized word sequences, but also semi-fixed and semi-free word sequences. Phraseology is argued in current paper to be the fundamental semantic unit because of the following 4 characteristics: (1) It covers more than half of all the tokens in actual discourse; (2) It has the function of *disambiguation*; (3) It displays co-selection of lexicons and structures; (4) it is the prefab in language. Phraseologies play the most important role in language communication. Collection of phraseologies can enhance the communicative efficiency of dictionaries. As a special form of corpus-driven method, phraseology-driven method is a recommended solution to the

problems related to concordance lines, and thus making corpus-driven methodology more efficient.

Currently, the identification and recognition of phraseologies is mainly based on statistical values. Although there are some defects in the current statistical identification algorithms, phraseologies can be effectively identified when the size of corpus is big enough.

Key Words: phraseology, phrase mining, Chinese corpus, corpus-driven, statistical value

I. Introduction

In most cases, people look up the dictionary for two purposes. One is to consult the meaning of the words for understanding, and the other is to look up collocations of specific words (not necessarily strange words) for use. The above two points are related to the definition and citation of the dictionary, which is the core content of dictionary compilation. With the help of the corpus, lexicographical definition and citation have overcome predicaments out of introspection, and revolutionary changes have taken place in both methods and means. The corpus displays the language material in use through concordance lines, and also provides context at the same time. However, dictionary compilation based on concordance lines has some obvious drawbacks. For example, it takes a lot of time to interpret the concordance lines. Especially when the data is too large, it is not feasible to read and analyze the concordance lines data carefully and thoroughly. As a result, a lot of valuable information is submerged in the concordance lines and is not available. Studies have shown that statistical method can be used to convert concordance line data into phrase-level linguistic data containing keywords and collocations. After the redundant information is filtered out, valuable information can be highlighted. Phraseology-driven method is not only as efficient as the concordance lines, but also saving time and energy. Phraseology-driven method can be considered to be the minimalist approach of corpus-driven methodology. (Li 2016: 40)

2. Phraseology in corpus linguistics

A phrase in a broad sense refers to a language unit consisting of two or more words. Interest of contemporary linguistics in phrases has been strong, so several similar terms have emerged, such as collocation, chunks, clusters, multi-word units, etc. Collocation has always been an important research area of corpus linguistics, and it is also one of the main topics for discussion in lexicography. But what's interesting is that there is no consensus on what is collocation in the linguistic community. The following different definitions reflect the differences in people's understanding of collocation:

- (1) Collocation is “the semantic compatibility of grammatically adjacent words”. (Hartman & James 2000: 22)
- (2) A term used in lexicology by some (especially Firthian) linguists to refer to the habitual co-occurrence of individual lexical items. (Crystal 2008: 86)
- (3) Collocation is the occurrence of two or more words within a short space of each other in a text. (Sinclair 1991: 170)
- (4) Collocation is the statistical tendency of words to co-occur. (Hunston 2002: 12)

The above definition presents us with a relatively confusing aspect of the collocation study. Collocations can refer to only the fixed structure, and can also include combinations of words with co-occurrence relations, regardless of whether the combination has independent meaning or not. Siepmann (2005: 409) thinks that collocation in the largest possible sense includes not only colligation but also phrases in general. At this point, the collocation has an all-encompassing nature.

Because of the large differences in understanding of collocation, corpus linguistics abandoned the term when creating a new field of research of phrases. A new term “phraseology” was coined to replace collocation and this new field of research is named “phraseology” using the same term.¹⁴ In 1998, the first book on phraseology entitled “Phraseology: Theory, Analysis and Applications” was published by Oxford University Press. Since then, the research on

¹⁴ Phraseology is promoted in corpus linguistics as a new field of research. Works in the field also take it as a countable noun, thus having “a phraseology”, “phraseologies”. In this article, “phrases” and “a phrase” are used instead of them in some cases in order to smooth the reading.

phraseology has gradually warmed up in the field of corpus linguistics and has gradually become one of the core research areas.

At present, the definition of phraseology has basically become uniform. It can be defined as: the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance. (Gries 2008: 6) Phraseologies are not only psychological co-occurrence of words, but more importantly, the frequency of co-occurrence between words that make up a phraseology is greater than theoretical frequency expected. This definition overcomes the purely subjective flaws in the judgment of phrases and is the basis for statistical recognition of phraseology.

The definition of Gries comprehensively summarizes the characteristics of phraseology. According to this definition, phraseology can be a phrase composed of two words, or a cluster of words composed of multiple words. Words in phraseology are not necessarily adjacent, and may be non-adjacent structural template like: *x(number) hours drive from y(place)*. Phraseology can be fixed phrases or special kind of free phrases. Fixed phrases include idioms, proverbs, allegorical sayings, special terms, etc. Free phrases refer to temporary structures observing grammatical rules, such as “red flowers, green leaves, structure of dictionaries, British universities”. The “structure of dictionaries” and “British university” are not phraseology, because the words that make up these strings of words are accidental co-occurrences, not statistically significant. They are random free phrases. While in contrast, words in “red flowers, green leaves” have a tendency to attract each other, and the co-occurrence frequency is also statistically significant. They are semi-free phrases and regarded as a form of phraseology.

Phraseology and corpus linguistics are closely linked to each other. In October 2005, 170 scholars from all over the world gathered in Louvain-la-Neuve, Belgium, to discuss researches on phraseology. The conference affirmed the contribution of corpus linguistics to phraseology, and the three books published after the conference strongly promoted the development of lexical studies around the world.

As Granger and Meunier (2008: xviii) put it, “Long regarded as a peripheral issue, phraseology is now taking centre stage in a wide range of fields, from natural language processing to foreign language teaching.” As a reference book that guides people to decode a language or encode in a language, the value of phraseology for dictionaries needs to be reconsidered.

3. Significance of phraseology in lexicography

In the fields of corpus linguistics, second language acquisition and natural language processing, the value of phraseology has received widespread attention, but in the field of lexicography in China, only a few scholars have noticed the value of it. (Xu 2013; Li 2014) In addition to traditional idioms, fixed collocations, etc., atypical phrases have not received much attention. For example, atypical phrases such as “谨慎乐观 (cautiously optimistic)”, “互利共赢 (mutual benefit and win-win)” and “小心台阶 (watch your steps)” in Chinese are likely to bring challenges to translators. The Chinese word “谨慎” have several synonymous translations, such as “careful, prudent, cautious”. But can they all be used as collocation with the English word “optimistic” to form an idiomatic expression? This is a question that the encoding dictionary needs to think about. For the decoding dictionary, the value of phraseology is also self-evident. For example, the meaning of “confidence man” in English is out of literal inference. If a dictionary fails to list the phrase, it will reduce the communicative value of the dictionary. In this section below, significance of phraseology will be discussed in detail from two aspects: language communication, dictionary compilation and research.

3.1 The significance of phraseology in verbal communication

3.1.1 Basic unit of meaning

If the unit of meaning is arranged in order, the order from small to large is sememe, word, phrase, clause, and text. Which one is the basic unit of meaning? The basic unit of meaning here refers to the language unit used by our intuition when using language to organize our thoughts. The basic unit of meaning needs to have indispensable features like modularity, high

frequency of use and unambiguity. Obviously, only words and phrases can be candidates of basic unit of meaning. Words are the smallest units of meaning that can be used independently, but words have a natural flaw: most words are ambiguous in meaning, such as the Chinese “打 (hit)” and the English “foot”. Without context, it’s hard to determine the part of speech of such words, let alone their meaning.

A phrase is a unit of meaning that is one level higher than a word. The important role of phrases in determining the meaning of words is affirmed by Firth. He (1957:12) said “You shall know a word by the company it keeps.” Sinclair has also repeatedly stressed that words are not isolated, they interact, and collocation is the key to the formation of meaning. (Moon 2008: 243) A phrase restores the minimal context of “language in use” in which the meaning of the word is manifested. Studies have shown that the two most important meanings of words, conceptual meaning and emotional meaning, can be clarified through the phrasal context (co-text) in which the word is placed. (Li 2016: 34-35)

Sinclair (2004: 36–37) advocates the idiom or phraseological principle and argues that language users follow a set of phrase rules in understanding and making sentences. There are a large number of preconstituted or semi-preconstituted blocks of language like finished building components that are stored in the minds of users, which play a critical role in encoding and decoding of language than words themselves. The language we use is largely a call to those pre-fabs.

According to statistics based on corpus, researchers found that phrases take a much larger part than words in encoding and decoding of language. Altenberg (1991) sampled LLC (London-Lund Corpus) and got to the conclusion that the proportion of phrases covers as high as 70% of all tokens in the corpus.

Because phraseology has features of modularity, high frequency of use, and unambiguity at the same time, we take them as the basic unit of meaning. They play a major role in verbal communication.

3.1.2 Co-selection of words and structure

Lexicon is not the only element for the formation of meaning in language, structure plays also an important role. Or in other words meaning of words move from opaqueness to transparency in structure. The focus on structural meaning can be traced back to Fries (1952). In *Structure of English*, he distinguishes between lexical meaning and structural meaning, and points out that to learn a language is to learn its structure made up of vocabulary. Harris believes that form and meaning (or grammatical structure and meaning) are indivisible, and his theory “Operator Grammar” proves that natural language is a “self-organizing system” through formal deduction. In this system, the structure and semantic attributes of words are clarified through the connection with other words. Harris also believes that our acquisition of structure is done through language contact. (Harris 1982, 1991)

The study of “structure” and “pattern” has spawned several linguistic theories such as “Construction Grammar”, “Pattern Grammar” and “Phraseology”. Construction refers to the pairing of form and meaning. It is the symbolic unit of language that is solidified in the mind of language users and shared by people in the same language community. Construction connects morphological, lexical and syntactic forms with semantic, pragmatic and textual functions. (Goldberg 1995, 2006) Pattern grammar is advocated by Hunston et al. Based on a large number of language examples and accumulated experience in the building of COBUILD and the following researches, they found that each word has a specific pattern to its own, in which the typical context of the given word is displayed. (Hunston & Francis 2000) Study of phraseology is proposed by corpus linguists. It not only emphasizes the computability of phrases, but also displays co-selection of vocabulary and structure, pattern and meaning. When we choose phrases for speech practice, we also choose vocabulary, grammar and pragmatic relations. (Partington et al. 2013: 26-30) It is just because phraseology combines lexical meaning with grammatical meaning that it has the advantage of eliminating ambiguity, because the meaning of phrases are usually transparent. From the English phrase “foot the bill” to idioms like “take a French leave” or templates like “{see} + [out of/from] the corner of [possessive]

eye” (Sinclair 2004: 171), the value of phraseology in meaning distinguishing and selection is immediate and significant. Phraseology integrates lexicon with grammar and make itself a self-sufficient semantic unit.

3.2 Significance of phraseology for lexicography

Because of the important role of phraseology in verbal communication, the value of it is self-evident for dictionaries since the purpose of them is for decoding and encoding of language.

First of all, phraseology embodies the communicative power of dictionaries. A dictionary is a reference book for intra-language or cross-language communication. The communicative power of a single word (or character) is weak in most languages, including Chinese. A single word cannot show the characteristics of “language in use”. For active bilingual dictionaries, interpretation of words does not have a reliable guiding effect on the encoding of language.

The Chinese character “打” is a polysemous word. Any interpretation of it in a bilingual dictionary will not be good enough for encoding. For the time being, let’s take one of its meanings for discussion. In the sense of “to hit an object with hand or instrument”, *New Century Chinese-English Dictionary* (2011) lists “strike; hit; knock; smash” as its English definition. If phraseology of “打” is neglected, it is helpless for translation of phrases such as “打翻(overturn) ; 打更(sound the night watches) ; 打鼓(play drum) ; 打屁股(spank)”. For bilingual dictionaries, the amount of phrases included is closely related to its communicative strength. The following is an example for language decoding. The *Collins COBUILD Dictionary* (2009) uses four sections for the English word *call*. The first three are sense categories of the word, and the last part is phrasal verbs. The dictionary lists more than 10 phrasal verbs related to *call*. The meaning of some phrases is difficult to infer literally, such as *call off*. Phrases listed not only affect the encoding power of the dictionary, but also have a positive correlation with its decoding power. For those phrases which are not self-explanatory, they will have a great impact on the communicative power of the dictionary if they are missing. The following is an example:

Meanwhile the defence ministry, which calls the shots on such vital questions as procurement and promotions, is staffed with career bureaucrats and political appointees.

In this sentence, the key to understanding is the phrase “call the shots”. Only when we know that the phrase means “make a decision, be in charge” can we understand this English sentence. *The Collins COBUILD Dictionary (2009)* does not list this phrase. It cannot be said that it is not a pity.

As a basic unit of meaning, phraseology is also the essential cognitive unit; in cross-language communication, it is the basic unit of translation. Therefore, whether it is a general language dictionary for natives to learn their mother tongue or a learner dictionary for second language learners, or a bilingual dictionary for translators, the listing of phrases is closely related to the communicative power of the dictionary.

Secondly, phraseology can be the starting point of corpus-driven definition for entry words in a dictionary. Phraseology-driven method can earn the best return. (Li 2016) Phraseology is effective in sense distinction. Since phraseology provides a co-text for “language in use”, the analysis and distinction of senses can be performed on the basis of a specific corpus composed of phrases. It is certainly feasible to write the definition of entry words or distinguish senses of polysemy with the help of concordance lines, but analysis of concordance lines usually takes too much time to be practicable. The inherent defect of analysis based on concordance lines is hard to overcome if the number of them is large enough. When concordance lines are further thinned into phrases, most redundant information is filtered and the key information will be highlighted. Phraseology-driven is an economical and efficient option for corpus-based definition and sense distinction.

Lastly, listing of phrases is cost-effective comparing to full-length sentences. Since definition offers little information for the use of language, full-length sentences are usually used in dictionaries, especially bilingual dictionaries to play a joint role with the definition. But full-

length sentences cover too much space. If phrases can guide readers to use the entry word properly in a given language, full-length sentences will not be necessary.

4. Phrase mining of Chinese based on statistical values

There are two basic methods for identifying phraseology: manual recognition and automatic recognition. From the current technical conditions, the accuracy of automatic recognition is lower than that of manual recognition. However, manual recognition is only suitable for small data, and phraseology identification based on a large corpus cannot but be done automatically. Automatic phraseology recognition is also a must-have function for corpus tools.

4.1 Basic method of automatic phraseology recognition

The automatic recognition of phrases relies mainly on statistical values. The simplest method of identifying phrases is based on the frequency of co-occurrences of node words and their collocates within a certain span. Wordsmith sets the threshold of frequency as 5. That is to say, if the frequency of co-occurrences of a node word and a collocate reaches 5 or more, it is considered to be a phrase. The following are the 10 most frequent collocates related to the node word “战争(war)” identified by Wordsmith (7.0) from CSWCM (Corpus of *Selected Works of Chairman Mao*):

Table 1: The top 10 collocates with “战争” sorted by simple frequency

Word	Set	Texts	Total	Total		L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
				Left	Right											
战争		1	951	79	79	23	24	23	7	2	793	2	7	23	24	23
的		1	803	333	470	50	60	49	48	126		195	62	82	53	78
是		1	255	119	136	16	21	28	37	17		41	27	19	21	28
和		1	133	63	70	11	6	10	17	19		27	4	12	15	12
中		1	121	32	89	3	2	2	23	2		52	16	11	9	1
在		1	112	84	28	16	3	6	27	32		6	5	2	6	9
不		1	79	36	43	8	13	10	5			7	11	7	7	11
了		1	71	43	28	7	6	11	7	12		2	5	3	10	8
革命		1	63	51	12	2	1		5	43		1	2	3	4	2
民族		1	46	37	9	3		4	17	13			3	2	3	1
也		1	44	12	32	3	2	6	1			4	6	9	8	5

As can be seen from Table 1, most collocates with “战争” are prepositions, conjunctions or other functional words. Their co-occurrence with node words has neither lexicographical nor statistical significance. They are not the phrases valuable for lexicography. The prominent co-occurrence frequency is due to their ultra-high frequency in the corpus (and the Chinese language as a whole), such as “的、是、和、在、了”. To overcome the shortcomings of simple frequency, linguists have devised some practical statistical methods for identifying phrases.

Evert (2004) has proposed more than 30 statistical algorithms. Wordsmith (Version 7.0) tool applies 7 methods for phraseology recognition, among which Z score, T score, and MI (Mutual Information) score are most commonly used. In addition, the Dice coefficient score is also an important method of identifying phrases. The Dice coefficient is between 0 and 1, and the larger the value, the stronger the collocability.

Using the above statistical methods, co-occurrences greater than the threshold (statistically significant) are highlighted, and most collocates with simple high frequencies are filtered. For example, taking “战争” as the node word, Z score highlights collocates “正规、正义、挑拨、国内、爆发、(战争)双方”. As the corpus capacity increases, more and more phrases will be recognized. Theoretically, if the corpus reaches a certain large scale, all important collocates with node words (such as “战争”) will be stored in it, and phraseology retrieval is nothing but a process of data mining process.

Table 2: Numbers of phraseology identified through 4 different statistical methods with reference to “dominant co-occurrence” value¹⁵

¹⁵ The “dominant co-occurrence” value herein refers to statistical score larger than the threshold value. For T score, it is larger than or equal to 1.645 (p value is 0.05); for Z score, it is larger than 2; and for MI score, it is larger than 3. In current research, the Dice coefficient larger than 0.03 is considered to be dominant.

	战争 (war)	人民 (peopl e)	革命 (revolutio n)
T score	389	255	359
MI score	530	371	434
Z score	211	152	183
Dice coefficient	96	84	105

It can be seen from Table 2 that co-occurrence word pairs based on MI score are most in number, those based on T score are closest to mean, and the co-occurrence words pairs based on the Dice coefficient are least in number.

Taking “革命” as the node word, out of the 30 top collocates identified through the above 4 statistical methods, 5 are the same, namely “社会主义” (socialism), “民主主义” (democracy), “动力” (motion), “对象”(object), “阶段”(phase). Collocates shared by T score, Z score and Dice coefficient are 14 out of the top 30 co-occurrence word pairs. Apart from the above 5, the other 9 collocates are “中国(China), 世界(world), 资产阶级(bourgeois), 无产阶级(proletariat), 三民主义(the Three People’s Principles), 任务(mission), 运动(movement), 特点(feature), 性质(nature)”. The ratio of consistency of the four methods is 17%; the ratio of consistency of T score, Z score and Dice coefficient is 47%.

T score and Dice coefficient fail to filter some function words and identify “的, 和, 了, 在, 也” as the top 30, while Z score and MI score do better. Collocates based on MI score are the most in number. The top 30 collocates of MI score show some discrepancy among the 4 statistical methods. The recognition accuracy of MI score is relatively poor.

Taking another word “战争(war)” as the node word, the data retrieved with the 4 statistical methods has the following characteristic:

- (1) Number of collocates identified in order: MI score>T Score>Z score>Dice coefficient.
- (2) Among the top 50 collocates identified by T score and Dice coefficient, more than half of them are function words or invalid to form collocations with the node word.
- (3) Among the top 50 collocates identified by Z score and MI score, most of them are lexical words and form significant collocations with the node word.
- (4) Z score and MI score show very high consistency among the top 50 collocates. The ratio of consistency is 64% (32/50). Most of them are valid collocates, such as “正规(regular) , 正义(just) , 掠夺(plunder) , 犬牙交错(jigsaw-like) , 流血(bloodshed) , 甲午(Jiawu)”.
- (5) Z score performs better than MI score, judging from the top 50 collocates. Apart from the shortcoming of over-identification, MI score fails to retrieve significant collocates as “残酷(cruel),爆发(break out)”, etc.

Result for “人民” (people) is similar to that of “战争”(war). Collocates recalled by MI score are the most in number; Most top collocates by T score and Dice coefficient are function words; Z score and MI score show consistency in identifying collocates. The following table is the top 20 collocates of Z score and MI score, highlighting collocates shared by the two.

Table 3: Top 20 collocates of Z score and MI score

Top 10 collocates of Z score	Top 10 collocates of MI score	Top 20 collocates of Z score	Top 20 collocates of MI score
全国	压榨	高压	自卫军
代表大会	呻吟	照旧	兴起
大众	乡	反帝	代表大会
压榨	区	县	县

呻吟	鼓动	总动员	总动员
乡	高压	广大	激起
区	照旧	利益	赞同
改良	冲破	隔离	踢
政府	衣食	尊重	障碍物
鼓动	他国	动员	深切

According to the data of the above three node words, we draw the following conclusion:

Different recognition algorithms have certain differences in accuracy and efficiency. T score is the least effective and it could be abandoned in actual phraseology recognition and extraction. Dice coefficient is unstable in identifying phrases, and the recognition result also has a lot of noise, especially for the top ranking collocates. The validity of MI score and Z score is better, but MI score has the disadvantage of over-identification. It's suggested to use Z score as the preferred method for recognition of phraseology and take MI score as a supplement.

4.2 Defects in the identification of phraseology based on statistical values

Identification of phraseology based on statistical values is currently the most effective method for automatic recognition of phrases, but the method has the following shortcomings:

(1) Defects of the algorithm itself. All algorithms have the problem of over-identification, among which the MI score is the most serious. For example, MI identifies “也罢, 关节, 游泳” etc. as the collocates of “战争”.

Contrary to over-identification is the overlooking and neglecting of significant phrases due to data sparsity. Even for MI score, words like “根据地(MI=2.68), 政治(MI=2.38), 党派(MI=2.36), 发生(MI=2.29), 民主(MI=2.25)” are not highlighted as collocates of “战争” because their statistical scores are less than the required threshold value.

The above drawback is a common problem of statistical recognition. The problem of data sparsity can be solved by enlarging the corpus, but over-matching is temporarily difficult to solve.

(2) Paradox of span. It is generally accepted that the well-suited span is 4 or 5, and the default value of Wordsmith is 5. But as the case for the Chinese language, the span of co-occurrence is not a fixed value. A smaller span will filter out a large number of phrases worthy of lexicographical attention, and a larger span will lead to excessive recognition of phraseology.

(3) Segmentation of Chinese. Compared with English, the automatic recognition of Chinese phrases based on corpus and statistics has the trouble of segmentation. Some problems caused by Chinese (automatic) segmentation are hard to solve so far.

Chinese word segmentation is currently based on a machine dictionary, usually using the maximum matching method or the maximum probability method. The algorithm of Chinese automatic word segmentation is far from perfect, and the resulting problems will also bring difficulties to the automatic recognition of phraseology. ICTCLAS, a popular tool for segmentation of Chinese, fails to identify “中非”(Sino-African) as a phrase and takes it as two

individual characters “中” and “非”. In doing so, phraseologies as “中非关系”(Sino-African relation), “中非合作”(Sino-African cooperation), etc. are ruled out.

5. Conclusion

The task of collecting phrases is usually undertaken by a dictionary (paper dictionary or machine dictionary), but in the field of lexicography, dictionary theorists and compilers have long been focused on relatively solid expressions. (Gries 2008: 3) Therefore, the task of recording phrases in the dictionary is far from complete, and a large number of phrases are intentionally or unintentionally excluded from the dictionary. Due to the importance of phraseology in language communication, active encoding dictionaries and machine dictionaries for language understanding or MT should take the responsibility to collect as many phrases as possible within their capacity.

The corpus-based phraseology recognition is mainly automatic, with statistical scores as the mirror of judgment. Human judgment is valuable only after the results are sorted out. Automatic recognition currently has some shortcomings, but the result of automatic phraseology recognition based on statistical scores is efficient and reliable in general. Overlooking of phraseology is a serious challenge to automatic identification, but as the volume and capacity of a corpus increases, problems due to data sparsity can become less serious. Over-matching is not a serious problem. In the final stage of artificial combing, mis-pairing or irrelevant phrases can be eliminated.

References

- Altenberg, B. (1991) ‘Amplifier collocations in spoken English’, in Johansson, S. and Stenstrom, A. B. (eds) *English Computer Corpora*, Berlin: Mouton de Gruyter.
- Crystal, D. (2008) *A Dictionary of Linguistics and Phonetics*, Oxford: Blackwell.
- Firth, J. R. (1957) *Papers in Linguistics 1934-1951*, London: Oxford University Press.

- Fries, C. C. (1952) *The structure of English*, New York: Harcourt, Brace and Co..
- Goldberg, A. E. (1995) *Constructions: A construction grammar approach to argument structure*, Chicago: University of Chicago Press.
- Goldberg, A. E. (2006) *Constructions at work: The nature of generalization in language*, Oxford: Oxford University Press.
- Gries, S. (2008) 'Phraseology and linguistic theory: A brief survey', in Granger, S. and Meunier, F. (eds) *Phraseology: An Interdisciplinary Perspective*, Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Harris, Z. (1982) *A grammar of English on mathematical principles*, New York: John Wiley.
- Harris, Z. (1991) *A theory of language and information: A mathematical approach*, Oxford: OUP.
- Hartmann, R. R. K. & James, G. (1998) *Dictionary of Lexicography*, London: Routledge.
- Hui, Yu (2011) *New Century Chinese-English Dictionary*, Beijing: Foreign Language Teaching & Research Press.
- Hunston, S. & Francis, G. (2000) *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*, Amsterdam / Philadelphia: John Benjamins.
- Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
- Li, Dejun (2014) 'Phraseological Unit (PU) and Its Statistical Identification', *Foreign Languages Research* 6: 8-13.
- Li, Dejun (2016) 'Corpus-driven Definition: Values, Drawbacks and Solution', *Lexicographical Studies* 2: 33-44+94.
- Moon, R. (2008) 'Dictionaries and collocation', in Granger S. and Meunier F. (eds) *Phraseology: An interdisciplinary perspective*, Amsterdam / Philadelphia: John Benjamins.

Partington, A., Duguid, A. & Taylor, C. (2013) *Patterns and meanings in discourse: theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam / Philadelphia: John Benjamins.

Siepmann, D. (2005) 'Collocation, Colligation and Encoding Dictionaries', *International Journal of Lexicography* 18: 409-443.

Sinclair, J. (1991) *Corpus Concordance Collocation*, Oxford: Oxford University Press.

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*, London: Routledge.

Sinclair, J. (2009) *The Collins COBUILD Dictionary*, Beijing: Higher Education Press.

Xu, Hai (2013) 'Phraseology and Collection and Arrangement of English Phrases in Dictionaries', *Lexicographical Studies* 5: 46-52+85+94.

TRILINGUAL ENGLISH-FILIPINO-PAMPANGO GLOSSARY OF CARPENTRY TERMS: BASIS FOR K-12 MATERIALS DESIGN

Eliezer V. David

Manila Tytana Colleges

Abstract

In general, the study aimed to develop a Trilingual English-Filipino-Pampango Lexicography of Carpentry Terms: Basis for K-12 Materials Design. Specifically, it aimed to identify the specific technical terms appropriate for carpentry lesson in K-12 then processes on how should these terms be translated using Newmark's communicative techniques in translation as well as the result of the evaluation of translated terms through Larson's Six Ways of Testing Translation. Finally, the study presented the terms in carpentry organized in a glossary form

This study is a qualitative type of research that used the descriptive method. The researcher collected English technical carpentry terms and definitions and translated the definitions to Filipino and Pampango as the target language applying Newmark's Communicative Techniques in Translation such. Later on, the two languages were processed to come up with a draft of an English-Filipino-Pampango Lexicography of Technical Carpentry Terms. The terms with definitions were arranged in a glossary format. The draft was tested using Larson's Six Ways of Testing Translation to have the final output English-Filipino-Pampango Lexicography of Technical Carpentry Terms.

Based on the study, the researcher found out that there were 220 technical terms appropriate for carpentry lesson in K-12. Technical carpentry definition of terms were translated using Newmark's communicative techniques in translation as cited in Evangelista (2012). There were three (3) terms rejected. Meanwhile, all Kapampangan translated definition of terms were accepted and twenty (20) definition of terms were modified as suggested by the validator. There are 217 translated definition of terms. These translated definition of terms were arranged alphabetically guided by Evangelista's (2012) study.

Based on the findings, it can be concluded that only 220 carpentry terms is acceptable in the study. Also, developing a trilingual lexicography needs to be done systematically through the use of methodologies and theories on lexicographic studies. Furthermore, validation of translated definition of technical terms can be done by a linguist or someone who possesses skills on translation procedures and people who are experts in the specific field where the glossary is intended for. Typical glossary is arranged in alphabetical order.

A glossary can be enhanced through continuous development. Therefore, the researcher recommended that the glossary could be expanded by adding more carpentry terms in verb forms. Also, translating other technical fields in Filipino or other dialects will further contribute in the enrichment and the development of Filipino lexicon. In addition, researchers who would prepare a similar glossary may administer questionnaires and rating scales to the target reader that will greatly contribute to the efficiency of the work and other researchers may further use other methods in arranging technical terms in a glossary form, if proven applicable.

Key Words: Lexicography, translation, technical terms

I. INTRODUCTION

The 1935 Philippine Constitution (article 13, section 2; Philippine Constitution 1935) referred to plans for “the development and adoption of a common language based on one of the existing native languages”. Commonwealth Act No. 570 (Commonwealth 1940) declared Tagalog as the basis of the national language, along with English. In 1959, Education Secretary Jose´ Romero issued a Department Order stating that the national language would be called Pilipino to distinguish it from its Tagalog base and to give it a national identity. The 1973 Constitution (Philippine Constitution 1973) designated Pilipino as the new national language and as an official language, along with Spanish and English. The 1987, post-People Power I Constitution (Philippine Constitution 1987) declared Filipino (now spelled with an F) as the national language as well as one of the official languages along with English. Spanish was dropped as an official language. The 1987 Constitution (in force as of 1994) also stipulated the creation of a new language body, Komisyon ngWikang Filipino (Commission of the Filipino Language) for the development and intellectualization of Filipino. Three Constitutions (1935, 1973 and 1987) have therefore decreed that the national language is Filipino; however, there seems a clear intent that English should remain as an official language.

The revised language policy of 1987 (Quisumbing 1987) prescribes the use of English for teaching maths, science and English while Filipino is prescribed for teaching all other subjects. However, observation shows that teachers typically begin teaching in the required language (either English or Filipino) and repeat the same content in the local language to ensure student comprehension of the curriculum content. Or teachers may code switch within the same statement (Gonzalez 1998; Young 2002). In practice, this means that local languages are used to explain the curriculum content to students rather than using those languages specifically as the media of instruction and teaching English and Filipino as subjects. In 2004, President Gloria Macapagal-Arroyo instituted a return to English as the primary medium of instruction in an effort to enhance the competitive edge of Filipinos in the international labor market (DepEd 2003).

With over 160 living languages, multilingualism is common in the Philippines and multilingual education has often been a controversial topic of conversation. Over the last several decades multiple studies, including PCSPE (1970), World Bank (1988), EDCOM (1991), PESS (1998), PCER (2000) and BESRA (2006), have recommended the use of the vernacular in the early years of education [Brigham and Castillo (1999)].

In an attempt to implement a mother tongue-based national educational program bridging learners from their first language to the languages of education, Filipino and English, Andrew Gonzalez, Secretary of the Department of Education, Culture and Sports (DECS) instituted DECS Memo No. 144 s. 1999 (DECS 1999) to develop foundational literacy skills (Cruz 2004). This program was expanded with the inclusion of more schools and more languages by DECS Undersecretary for Programs and Projects, Isagani Cruz (DECS Memo No. 243 s. 2000, DECS 2000). This resulted in the expansion of the 1974 Bilingual Education Policy to a “still-unnamed and unacknowledged Multilingual Education Policy” (Cruz 2004). The Basic Education Curriculum (DECS, changed to DepEd in 2001, Order No. 43s 2002, DepEd 2002) as implemented by Secretary Raul Roco maintained a focus on the central role of language in education and retains the multilingual policy begun in the expansion of the Regional Lingua Franca Program. The most recent and probably strongest statement is DepEd Order No. 74 2009 (DepEd 2009), Institutionalizing Mother Tongue-Based Multilingual Education (MLE). At the time of writing it is the intent of the Department of Education to implement fully a strong MLE programme beginning in each learner’s first language and bridging them to Filipino and English.

Language plays an important role in the society. It manifests the common system that mirrors the social, political and technological improvement of its people. Vocabulary of the society progresses with so many complications and difficulty due to its practically and technology influence. Evangelista (2012) stated that a development of a society is based on a common vocabulary used. Source scientist undiscovered the importance of native language as a means of eliminating language barriers in the society. Fishman (1986),

a sociologist, presented out those studies on native language will minimize ignorance of ethnicity. A matter that according to him, is greatly in need of an unhurried review as a social parameter. Another sociolinguist, Rosal (2011) added that there is a recovering need to undertake more researches aims towards the development and preservation of the mother tongue.

One of the efficient and useful devices in attaining the goal and solve this predicament is through the use of a dictionary or a bilingual specialized glossary for technical and scientific terminologies.

Theoretical Framework

Peter Newmark took the lead in making translation a craft in its own right. He made translation theory in remarkable ways, describing the conversion of a text from one language to another as both a science and an art. He proposed communicative translation that attempts to produce on its readers an effect as close as possible to that obtained on the readers of the original. It is likely to be smoother, simpler, clearer, more direct, more conventional, conforming to a particular register of a language, tending to undertranslate, to use more generic, hold-all terms in difficult passages for the comprehension and the response of the target language receptors. Moreover, he proposed techniques to use for communicative translation in attaining this goal in translation. These techniques are: (1) transposition, (2) cultural equivalent, (3) functional equivalent, (4) reduction, (5) modulation, (6) addition or expansion, (7) naturalization, (8) adoption or transference, (9) lexical synonymy, (10) one-to-one translation, (11) componential analysis, (12) thorough translation, (13) descriptive equivalent or amplification, (14) recognized translation, (15) compensation, (16) paraphrase, (17) improvements, and (18) couplet. For him, the translator acquires a technique in which the process to be followed takes into account of comprehension, interpretation, formulation and recreation. (Newmark, 1982)

Mildred Larson, an international translation consultant and proponent of meaning-based theory, explained that a translator needs to be sure if the translation is accurate, clear, and natural. These three features are important throughout the translation, so the entire translation must be checked for each one. According to her, the ideal translation will be accurate as to meaning and natural as to the receptor language forms used. The intended audience of the translation who is unfamiliar with the source text will readily understand it. Furthermore, she proposed several ways of testing translation that will help the translator to ensure the accuracy, clarity, and naturalness of the translated work. These ways are: (1) comparison with the source language, (2) back-translation into the source language, (3) comprehension checks, (4) naturalness and readability testing, and finally, (5) consistency checks. These translation testing needs to be done systematically, and notes need to be taken carefully as these notes are important, not just for improving the

translation which was checked, but also for evaluating the errors which are repeated again and again. (Larson, 1984)

A glossary as conceptualized in their study is a collection of terms limited to a special area of knowledge and usage (Juano, 1980). In addition, the terms explorations with their meanings are expressed in the same or another language.

In this study, the meanings are expressed in Trilingual: English, Filipino and Pampango.

Statement of the Problem

In general, the study aimed to develop

Specifically, it aimed to answer the following:

1. What are the specific technical terms appropriate for carpentry lesson in K-12?
2. How to translate the terms using Newmark's communicative techniques in translation?
3. What is the result of the evaluation of translated terms through Larson's Six Ways of Testing Translation?
4. How were the terms in carpentry organized in a glossary form?

II. METHODS AND PROCEDURES

Research Method

This study is a qualitative type of research that will use descriptive method. This method will be designed to gather information about the present existing conditions. This method will be used to describe the native and situation, as it exists at the time of the study (Travis, 198). Description research involves collection of data in order to answer questions concerning the current status of the subject of the study (Evangelista, 2012).

Respondents of the Study

The respondents of the study are speakers of Filipino and Pampango languages, selected people in the field of carpentry and translation and people practicing teaching carpentry or languages for at least two years. They are the ones who are knowledgeable enough to answer the problems posed in the present study. They answered the questionnaire that the researchers gave them which supplies the information the researchers need.

Instrument and Techniques

Unstructured Interview

The researcher will use an unstructured interview in getting personal data such as name, degree, position (if employed), years in service and field of specialization. These varied information will reveal the capability of the validators. Moreover, the researcher will formulate specific criteria in choosing the validators of the study. These are

1. The validator is a speaker of English and Filipino and/or Filipino and Pampango languages;
2. The validator is in the field of carpentry and/or translation;
3. The validator is practicing the profession for at least two years.

Validation of the Instrument

In the validation proper, validators were given a copy of translation. Here, the validators put their evaluation through putting a check mark on the blank space provided if they accept, reject, or modify the translation made by the researcher.

Data Gathering Procedure

Phase 1: Data Gathering of Carpentry and Their Definition

In the compilation of terms, the researcher used available books, journals, magazines, encyclopedias, dictionaries and world wide web resources on carpentry. From these varied materials, the researcher gathered and collected various carpentry terms and asked a civil engineer who is also teaching carpentry in a technical school to identify the terms that would suit per lesson in the K-12 module.

Phase 2: Preliminary Translation by the Researcher

The researcher being a Trilingual English-Filipino-Pampango speaker translated English as the target language through the gathered carpentry terms and definitions into Filipino and then Pampango using Newmark's Communicative Techniques in Translation. These translations were written beside their English counterparts in the index cards. In this translation process, the researcher was guided by the studies and concept in relation closely related to this work.

After the various carpentry terms and definitions had been translated, the cards were arranged alphabetically and transferred on papers.

Phase 3: Testing the Translation

When the index card entries had been transferred on paper, the translation was tested using Larson's (1984) Six Ways of Testing Translation. The following ways of testing a translation were applied.

3.1 Comparisons with the source language

This was done in order to check the content of every entry. This check was applied to ensure if all the information was included. The comparison check was made by having a draft of the translation encoded with double space and wide margins on both sides so that ideas can be written in the margins and the alternatives can be written above for later evaluation. Checking with the source language guaranteed that the meaning and dynamics of the source text was completely communicated to the target users of the trilingual glossary. Since the comparison test is a self-check process, the test was done by the researcher himself.

3.2 Back translation

This was done by another Trilingual English-Filipino-Pampango speaker who had not read the source text used by the researcher. The Trilingual English-Filipino-Pampango speaker was given a copy of the translated terms in Filipino and asked to write out the meaning he got from it back into source language. This back translation allowed the researcher to know if the translation was being completely communicated to this person.

3.3 Comprehension test

This test was designed to find out whether or not the translation is understood correctly by the speaker of the target language; two speakers participated in this test. With a copy of the glossary in hand, the researcher asked these persons to say something about the meaning of the translation. Comments were noted on the sides of the glossary for further evaluation and modification.

3.4 Naturalness test

This test aimed to find out if the form of translation is natural. Language and carpentry experts were asked to participate in this test. They reviewed the entire glossary and asked to make comments on the margins as they read along.

3.5 Readability test

This test was done by the translator and a speaker who is also Trilingual in English and Filipino and/or Filipino and Pampango. The reader was requested to read the translation aloud. As he read, the translator gave attention to the place or places where the reader hesitated; also, when he stopped and re-read the translation. All these will be noted by the researcher for modification of the translation.

3.6 Consistency check

This check has to do with the content of the translation and the technical details of the presentation. In this work, aside from the researcher, a language professional was requested to go over the entire glossary and check the consistency of the spelling, capitals, and punctuation of the translated text. A thorough and careful proof-reading of the Trilingual glossary will be made.

Phase 4: Revision of Glossary

When all the various tests had been carried out, results were analysed by the researcher and changes were applied to the Trilingual glossary. After the translation had been reviewed and improved, the researcher prepared the second draft of the glossary.

Phase 5: Evaluation and Validation of Glossary

Qualified evaluators were asked to evaluate the Trilingual glossary by reacting to the translation given through an evaluation instrument. If they accepted the translation, they had to put a check mark on the blank space on the right for “Accept”. If they rejected, they had to put a check mark on the blank space for “Reject”. If they wished to modify the translation, they had to put a check mark on the space for “Modify” and then write their modification to the translation on the blank space below the original translation.

Phase 6: Writing the Final Copy of Trilingual English-Filipino-Pampango Lexicography of Carpentry Terms.

After the evaluation of the experts, the researcher reviewed the Trilingual glossary for modification, addition or changes to be made. This careful check was made to produce more polished Trilingual glossary of Carpentry.

The Trilingual glossary of Carpentry contains entries that are single-word, compound or multi-word. The entries were defined either in single-word definition used in Filipino equivalent or multi-word or phrase definition of the term. The English definition or equivalent of the entry word is given first. The followed by the Filipino equivalent or translation arranged alphabetically. Each Carpentry entry is composed of the following:

1. Lexical Entry
2. Translated Entry
3. Definition

III. PRESENTATION OF RESULTS AND DISCUSSION

A. Specific Technical Terms Appropriate for Carpentry Lesson in K-12

The researcher prepared a checklist purposively for an engineer who is practicing his expertise both in the field and academe. A checklist is a device which contains the items to be observed and space for number or check marks or short verbal entries. It contained English terms used in carpentry gathered by the researcher. The terms were compiled from books, journals, magazines, encyclopedias, dictionaries and world wide web resources on carpentry. After which, the engineer was asked to designate each term on specific lessons available in the K to 12 Basic Education Curriculum for Technology and Livelihood Education learning module for carpentry. The listed terms given were arranged alphabetically. Also, another sheet was provided which includes the K-12 curriculum guide for carpentry program. The curriculum guide is divided into five (5) lessons. To identify the appropriateness of each terms, the engineer to answer the checklist was asked to number the terms 1 – 5 depending on what lesson will match the term be encountered. In cases like term/s suit to two or more lessons, the engineer just placed more the one number to the item. After the list of carpentry terms, a space at the bottom part was provided for the engineer to write down additional equivalent words/terms that they know.

After terms were being assigned, the researcher summarized the terms by plotting a column for the lesson as well as the learning outcomes and adjacent to that were the terms that are applicable per lesson.

B. Translation of the Technical Carpentry Terms using Newmark's Communicative Techniques in Translation

Each technical carpentry term is already considered a lexical entry. The researcher used English as the source language and identified Filipino as 1st target language (TL1) and Pampango as the 2nd target language (TL2). Each lexical entry in the source language was written in index cards, before the corresponding translation to its target language was made employing Newmark's Communicative Techniques in translation. These techniques are namely as follows:

- a. Transposition – the shift in sentence construction;
- b. cultural equivalent- the use of vocabulary words that has appropriate cultural meaning from the source language to the target language;
- c. functional equivalent- the use of idioms or figurative language in translation based on function and meaning;
- d. reduction or contraction – the shortening of form but ensuring the same context;

- e. naturalization, adoption or transference- adapting or borrowing words not found in the culture of the target language;
- f. one-to-one translation- the transfer word for word or in verbatim;
- g. componential analysis- dividing a sentence by segment;
- h. thorough translation, descriptive equivalent or amplification- translating collocations in the same manner as the source language to the target language
- i. recognized translation- translating a word by giving a description to the term

Evaluation of Translation through Larson's Six Ways of Testing Translation

The

researcher adapted Larson's six ways of testing the translation in this phase of the study. Larson's Six Ways of Testing Translation comprises comparison with the source language, back translation, comparison test, naturalness test, readability test and consistency check with the help of the experts like engineers, civil engineering professors, language professors and translators to have the final output which is a trilingual glossary of English-Filipino-Pampango Glossary of Technical Carpentry Terms:

- A. Comparison with the Source Language- A careful comparison with the source language was made several times to ensure that all the necessary information about each term was included, specifically in getting the definition of each and every lexical entry through the various references like books, carpentry journals and world wide web. Since, this is a self-check process; this phase was made by the researcher.
- B. Comprehension Test- The researcher wanted to know if the translated terms were understood correctly by the speakers of the target language. So, this test was carried out with two validators. The researcher gave a copy of the glossary in hand and the validators were asked if they comprehended the meaning of the translated terms. The comments were noted for further modification of the glossary. In this test, three (3) translated technical terms from the first target language (TL1) – Filipino were rejected. These are bollard, dado joint, and fascia. These terms were removed by the researcher in the final copy the glossary. Meanwhile, eleven (11) translated technical terms from the second target language (TL2) – Pampango were modified based on the recommendation of the validator.
- C. Naturalness Test- The aim of this test is to find out if the form of translation is natural. The validators were given a copy of the glossary and asked to make comments on the content of the translated terms. In this test, twenty (20) translated technical terms from TL1 were modified while nine (9) translated technical terms from TL 2 were modified based on the recommendation of the validator.

- D. **Readability Test-** This test was done by the translator and a speaker who is fluent in English, Filipino and Pampango. The speaker who is fluent in English, Filipino and Pampango was tasked to be the reader. The reader was requested to read the translation aloud. As she read, the researcher focused on the way the trilingual speaker read the translated terms. Apparently, it sounded clearly to the researcher that showed clearness and clarity of the translated terms. At this phase, no revision is necessary based on the readability test.
- E. **Consistency Check-** This check had to do with the content of the translation and the technical details of the presentation. In this work, aside from the researcher, a technical writing professor was requested to go over the entire glossary and check the consistency of the spelling, capitals, and punctuations of the translated terms to make a thorough and careful proof-reading of the trilingual glossary.

Validation of Glossary

After all the various tests had been carried out, results were analyzed by the researcher and changes were applied to the bilingual glossary. The second draft was made and it was given to three qualified validators. These three validators were given an evaluation instrument and asked to put a check mark on the blank space on the right for “Accept”. If they reject, they have to put a check mark on the blank space for “Reject”. If they wish to modify the translation, they have to put a check mark on the space for “Modify” and then write their modification to the translation on the space provided or may have the option to overwrite corrections.

Final Copy of the Trilingual English-Filipino-Pampango Carpentry Terms Arranged in a Glossary Form

After series of tests with the validators, a total of Three (3) translated definition terms from TL1 were omitted because these terms were rejected during the comprehension test and there were twenty (20) terms recommended to be modified by the validator during the naturalness test. While, Eleven (11) translated definition terms from TL 2 were modified by the validator during the comprehension test while a total of nine (9) words were asked to be modified during the naturalness test.

A total of 217 translated definition of English-Filipino-Pampango carpentry terms were chosen to be in the final copy of the glossary.

After the experts made the validation, the researcher reviewed the trilingual lexicography for modifications, additions, and changes. This careful and thorough check made to produce more polished trilingual terms in carpentry.

There are 217 translated definition of terms. These translated definition of terms were arranged alphabetically guided by Evangelista's (2012) study.

The trilingual lexicography on carpentry contains 217 entries that are single-word, compound or multi-word. These entries were defined either in single-word definition used in the Filipino equivalent or multi-word or phrase definition of the term. Researcher need not to include the word class for it has been mentioned that the dictionary will only be limited to nouns. The English definition or equivalent of the entry word was given first. Then, followed by the Filipino equivalent or translation then Pampango which are arranged in alphabetical order.

IV. CONCLUSIONS

This research aimed to translate English terms in carpentry to Filipino and Pampango as the target language. In addition, Newmark's communicative techniques in translation were employed and Larson's testing translation was used in evaluating the translated terms.

It can be concluded that From the 300 technical carpentry terms gathered from different references such as books and journals, only 220 terms were appropriate for carpentry lesson in K-12. Trilingual glossary in this case is critical because not all terminologies found needed in the field of carpentry is needed when placed in a glossary format.

Developing a trilingual lexicography on technical fields like in criminal law, garment technology, farming and midwifery is an indispensable task that needs to be done systematically through the use of methodologies and theories on lexicographic studies. It is a great advantage that works in translation and lexicography have been published and open for public reference which leads to a guidance in properly producing a reliable material for language learning and development.

Validation of translated definition of technical terms could be done by a linguist or someone who possesses skills on translation procedures and people who are experts in the specific field where the glossary is intended for.

Typical glossary is arranged in alphabetical order as presented in Canlas (2010), Rosal (2011), Delos Reyes (2012) and Evangelista (2012).

This paper has implications to technical pedagogy as it allows learners to relate and even analyze terminologies that can be utilized in the field. It also facilitates awareness and appreciation of the holistic role of language amidst the challenges in the K12 curriculum.

A translation work can be enhanced through continues development. The researcher recognized that there are possible insufficiencies. The produced trilingual lexicography of carpentry terms invite for a more empirical and positivist effort to test its universality across a more aggregate linguistic population.

IV. References

- [1] Ballena, C (2005). *Development and Validation of an English-Filipino Dictionary of Philosophy*. Unpublished doctoral dissertation, Philippine Normal University, Manila.
- [2] Belza, Elgin I. 1999.” The Monophology of Binukid Verbs.” Unpublished Special Project. Philippine Normal University
- [3] Canlas, R. (2010). Development and Validation of a Kapampangan – English glossary of Selected Kapampangan Idioms. Unpublished master’s Thesis. Philippine Normal University.
- [4] Casanova, Arthur P. (1998). Development and Evaluation of Monolingual Filipino Dictionary on Drama and Theater Arts. Unpublished Doctoral Dissertation, Philippine Normal University.
- [5] Del Rosario, Fe Laura, (1973). *A Model for an Etymological- Monolingual Dictionary of Tagalog*. Unpublished master’s thesis. Ateneo de Manila University. Unpublished Master’s Thesis. Philippine Normal University.
- [6] Delos Reyes, Riza I.(2012). *Development and Validation of English-Filipino Glossary of Midwifery*. Unpublished Master’s Thesis. Philippine Normal University.
- [7] Evangelista, J. (2012). *Bilingual English-Filipino Glossray of Legal Terms*. Unpublished Master’s Thesis. Philippine Normal University.
- [8] Evina, Julie Hope Timotea P. 2013. *English-Tagbanua Glossary of Agricultural Terms*. Unpublished Master’s Thesis. Philippine Normal University.
- [9] Larson, M. (1997). *Meaning-Based Translation: A Guide to Cross-Language Equivalence*, 2nd Edition. University Press of America
- [10] Tomášek, M. (1990). *On Selected Problems in Translation of the Legal Language*. Transtologica Pragensia IV, Acta Universitatis Carolinae Philologica. Prague: Karolinum
- [11] Zgusta, L. (1971). *Manual of Lexicography*. Paris: Academic Publishing House of Czechoslavac Academy of Sciences.
- [12] J. P. Wilkinson, “Nonlinear resonant circuit devices,” U.S. Patent 3 624 12, July 16, 1990.
- [13] R. J. Vidmar. (August 1992). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21(3). pp. 876-880.

Author’s Biography



Eliezer V. David was born in Manila, Philippines on May 6, 1988. He is an English Instructor of the College of Arts and Sciences-Languages Department Manila Tytana Colleges. He is also the subject area coordinator for the English Unit of the SHS Department. He obtained his bachelor's degree at Philippine Normal University with the degree of Bachelor in Secondary Education Major in Speech and Theater Arts where he also completed academic units for Master of Arts in Linguistics. He finished Master of Arts in Teaching English as a Second Language at Manuel L. Quezon University where he also earned units in Doctor of Philosophy in English. The author has taught in different private, local and state colleges in Manila. He also has extensive background and training experience on Linguistics, Speech Communication and Theater Arts.

DO AFRICAN LANGUAGE CHILDREN'S DICTIONARIES MEET THE NEEDS OF THEIR TARGET USERS?

Elsabé Taljard and D. J. Prinsloo

Department of African Languages

University of Pretoria

Abstract

Judging by the dearth of research on children's dictionaries, lexicographers do not seem to regard this dictionary genre as worthy of serious academic attention. The situation is even worse for children's dictionaries published in the (South) African languages. With the exception of the publication of Gouws, Prinsloo and Dlali (2014), children's dictionaries have not been the focus of lexicographic research, attracting little more than a sidelong reference in general discussions on lexicographic theory and practice. In this paper, we investigate two aspects of children's dictionaries, i.e. the initial conceptualization of the dictionary, and secondly, the extent to which selected children's dictionaries meet with the user's expectations. We focus on three children's dictionaries, i.e. Series of *Foundation Phase dictionaries* (FPD), 2010, published by Maskew Miller Longman; *The Official Foundation Phase CAPS dictionaries* (OFPCD) 2018, published by the South African National Lexicography Units and *The Jul'hoan Children's Picture Dictionary* (JCPD), 2014, published by University of KwaZulu-Natal Press. We argue that dictionaries for African language speaking children must be conceptualized from an African (language) perspective. We indicate that the mismatch between conceptual relationships and linguistic form and function is the result of initial conceptualization of the dictionary from an European point of view. With regard to user's expectations, we refer specifically to the need to take the target user's frame of reference into consideration with regard to selection of lemmas and illustrative material. We argue that moving too far beyond users' frame of reference leads to an overload of lexicographic information. Our study uses the user's perspective as a starting point, but we also refer to Beyer's (2014) theory of lexicographic communication.

Key Words: children's dictionaries, African language dictionaries, user's perspective, theory of lexicographic communication

Introduction

Considering the fact that dictionaries for children are the gateway to the establishment of a dictionary culture and may thus be instrumental in the process of life long learning, it is surprising that the compilation and evaluation of such dictionaries receive so little attention from researchers in lexicography. Within the South African context, dictionaries aimed at target users in the first three years of formal schooling, i.e. the so-called foundation phase, fulfil a crucial function in familiarizing young users with the dictionary as utility tool. Within lexicographic theory, provision must be made for the planning of such dictionaries, since they are no less important than for instance extensive explanatory monolingual dictionaries, which in a certain sense represent the upper end of lexicographic endeavour. In this regard, Gouws (2013) remarks that “a general theory of lexicography [...] should assist any lexicographer planning any new dictionary, also foundation phase dictionaries”. In this article we reflect on the design of three dictionaries / dictionary series, all presumably aimed at first time dictionary users, aged approximately 7 to 9 years of age. We will focus on the initial conceptualization of the dictionary, and secondly, discuss the implications of a truly Afrocentric approach to compilation of dictionaries for African language children. Our study uses the user’s perspective as point of departure, but we also refer to Beyer’s theory of lexicographical communication (Beyer, 2014). In our discussion we will refer to the following dictionaries:

- A series of Foundation Phase dictionaries, 2010, published by Maskew Miller Longman (henceforth MML dictionaries);
- The Official Foundation Phase CAPS dictionaries, 2018, published by the South African National Lexicography Units (henceforth NLU dictionaries); and
- The Jul’hoan Children’s Picture Dictionary, 2014, published by University of KwaZulu-Natal Press (henceforth JCP dictionary).

The first two titles referred to both constitute a series of bilingual dictionaries, each dictionary treating English and an African language, whereas the third dictionary is free standing, treating three languages, e.g. Jul’hoan, Afrikaans and English. Jul’hoan is a Northern Khoesan language, spoken by San people in Namibia and Botswana. It is an endangered language, spoken by 11 000 speakers.

Conceptualization of the dictionary: top down or bottom up?

It is generally accepted within the lexicographic community that the compilation of any dictionary must be preceded by proper planning, cf. Gouws and Prinsloo (2005: 13-19). During the planning phase, issues such as the organisation plan, the genuine purpose and lexicographic function and the conceptualization of the dictionary are considered. The identification and formulation of the genuine purpose of the dictionary must precede all other lexicographic processes, since these processes will be

driven by the genuine purpose. With regard to the MML and NLU dictionaries, their genuine purpose is stated as assisting users with text reception, text production as well as serving their cognitive needs. In addition to the explicit formulation of the genuine purpose of these dictionaries, the lexicographers have stated a complementary aim, i.e. the establishment and promotion of a dictionary culture, familiarizing target users with the dictionary as a practical instrument that can play an important role in the process of life-long learning, cf. Gouws et al (2014: 27).

Being school dictionaries, it is to be expected that the content of the MML and NLU dictionary series must be aligned and integrated with the curriculum for the relevant educational phase at which these dictionaries are aimed. Both the MML and NLU dictionaries emphasize the fact that the content is based on information provided in the Curriculum Assessment Policy Statement (CAPS) documents. These documents are official documents, emanating from the Department of Basic Education, which stipulate the policy on curriculum content and assessment for each educational phase. In the users' guide to both dictionaries, it is indicated that the dictionary is theme based, and that the selection of themes is done according to the outlines of the CAPS for Foundation phase.

Basing a dictionary and its contents on a prescribed school curriculum implies that the conceptualization of the dictionary is in essence a top down approach. The contents of the dictionary, i.e. the themes to be addressed, as well as the lemma list are to a certain extent pre-determined. In the user's guide to the NLU dictionaries, it is stated that the themes include the basic words of objects and actions learners may encounter in everyday life. This statement raises an issue that is particularly pertinent to children's dictionaries, i.e. the notion that lexicographers should take cognisance of the user's frame of reference (FoR). The other side of the lexicographic coin is that lexicographers also operate from within their own frame of reference (FoR). In the case of children's dictionaries it is almost inevitable that the (adult) lexicographer's frame of reference and that of the child as target user will not coincide, resulting in an unsuccessful transfer of the lexicographic message to the user. Determining the FoR of the target user should form part of the dictionary planning process and should not be left to supposition. One way of doing this would be to involve the target user in the planning and conceptualization of the dictionary, in as far as this is feasible. Such an approach would constitute a bottom-up process, which could contribute to bridging the gap between the perceived FoR of children and their actual FoR. In this regard, the planning and conceptualization of the JPC dictionaries represent an innovative and creative approach. The planners of this dictionary had the advantage of a small, homogeneous body of target users, whose FoR is relatively easy to determine. In an information pamphlet accompanying the dictionary, it is stated that members of the community lead the way in the selection of themes, lexical entries, design and layout. The themes that have been selected use the real life experience(s) of the target user as point of departure and reflect the lived experiences of the target users. These themes are:

animals, birds, insects, reptiles and creepy crawlies, home and the family, hunt, gather and dance. Selecting themes that children as dictionary users can relate to and which fall within the ambit of their FoR can contribute not only to a successful dictionary consultation process, but also to the establishment of a dictionary culture.

Eurocentric dictionaries for African language children

A children's dictionary that is truly Afrocentric is much more than a dictionary characterized by a superficial adaptation of illustrations depicting black children instead of their white counterparts. It is a dictionary that is in the first instance sensitive to portraying anything European as the default and as an ideal that has to be emulated. It has to make provision for the fact that the concept 'house', for example, does not have to be portrayed as a typical European house, but that it can be a much less formal dwelling, which in turn would take the user's FoR into consideration. An Afrocentric dictionary should furthermore be sensitive to what can at best loosely be termed African values. It has to take cognisance of the fact that the concept 'my family' in many African households are single parent or child headed families, or that it may include the so-called extended families. Lastly, dictionaries for African children should be much more than dubbed over versions of dictionaries initially conceptualized for English (or Afrikaans) speaking target users. It is all too often the case that a dictionary conceptualized for English (or Afrikaans) is used as a blueprint for an African language dictionary, or a bi- or multilingual dictionary containing an African language as one of its language pairs. Such an approach does not always work to the advantage of African language dictionaries in that it can result in a conceptual mismatch between conceptual relationships and (linguistic) form and function. This is especially relevant for children's dictionaries, where emphasis is on the establishment of conceptual relationships. Adjectives are a case in point. Words indicating attributes are mostly adjectives in English. Therefore, apart from the conceptual relationship between these words, there is also a functional and syntactic relationship between words such as 'big', 'small', 'bright', 'lazy': they can be used attributively, e.g. 'lazy teacher' or predicatively, e.g. 'the teacher is lazy', they have degrees of comparison, etc. English speaking dictionary users therefore have the advantage of a link between syntax, function and conceptual relatedness. Seen from the viewpoint of the African languages, this link is absent. Attributes in African languages are expressed by means of a number of different constructions – some attributes in Sepedi for example, are expressed by means of adjectives, some by means of so-called possessive constructions or nominal relatives, and others by means of verbal relative constructions. Since adjectives are a closed class in African languages, it would make sense to group members of this category together in one dictionary theme, e.g. 'Describing things', so that the functional and conceptual relatedness, i.e. indication of attributes and syntactic properties is aligned.

Conclusion

In order to meet the needs of their target users, African language dictionaries for children need to be conceptualized from an African (language) perspective. Such an approach will ensure a match between conceptual relatedness and syntactic form and function. By taking cognisance of their target users' FoR, lexicographers can produce dictionaries which truly fulfil the genuine purpose of children's dictionaries. A bottom up approach to dictionary compilation in which the target user is actively involved in the compilation process, is not only desirable, but also feasible, as illustrated by the JPC dictionaries.

References

Dictionaries

Jones, Kerry L. and T.F. Cwi (eds.). 2014. *Jul'hoan Tsumkwe Dialect/Prentewoordeboek vir kinders/Children's picture dictionary*. Pietermaritzburg: University of KwaZulu-Natal Press.

Mabule, M. 2010. *Longman pukuntšu ya sehlopaqase: Sepedi/English*. Cape Town: Maskew

Miller Longman.

The Official Foundation Phase CAPS Sesotho sa Leboa – English picture dictionary. Pretoria: Sesotho sa Leboa National Lexicography Unit.

Other sources

Beyer, H. 2014. Explaining Dysfunctional Effects of Lexicographical Communication. *Lexikos* 24, 36 – 74.

Gouws, R.H. 2013. Establishing and developing a dictionary culture for specialised lexicography. In Jesenšek, V (ed). *Specialised lexicography Print and Digital, Specialised dictionaries, Databases*. Lexicographica Series Maior 44, 51 – 62.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and practice of South African Lexicography*. Stellenbosch: SUN PReSS, AFRICAN SUN MeDIA.

Gouws, R.H., D.J. Prinsloo and M. Dlali. 2014. A Series of Foundation Phase Dictionaries for a Multilingual Environment. *Stellenbosch Papers in Linguistics* 43: 23 – 43.

INTERACTIVE TERM DEFINING MODULE: A MODAL OF LEXICOGRAPHY TERMS DICTIONARY

Professor Doctor Erdoğan BOZ

Eskişehir Osmangazi University

Assistant Professor Bülent ÖZKAN

Mersin University

PhD. Nilay GİRİŞEN

Anadolu University

Abstract

The purpose of the project is to interactively define the terms of Turkish Lexicography, which are listed in the Lexicology Module of Do it Yourself Corpora (DIY Corpora), by using the Terminology Module of DIY Corpora. In this sense, the title of the project is “Interactive Term Defining Module” (ITDM). This project aims to comply with the International Standards Institute (ISO) norms and the criteria of definition in the terminology and to create a module that will enable the definition to be made interactively by multiple authors, especially for the new terminological dictionaries. In line with the method implemented, the expected result of this project is to develop an interactive term defining module in accordance with ISO standards which are open to multiple authors and editors by using computer technology. Although there are some definition modules in the literature, it is seen that they are not interactive term defining modules that comply with ISO standards, also the definitions are made in a traditional way and they are not consistent with themselves. The project team and the staff have an interdisciplinary qualification in line with the aims and objectives of the project. The project team will include researchers and staff from different disciplines such as linguists, lexicology and terminology experts. In addition, the proposed project foresees the participation of all stakeholders in the ITDM creation process such as field instructors, students and linguistics / lexicography specialists. In this study, the method of creating an interactive term defining module will be presented. In addition, we are studying on this project as the second step of the project named “A Corpus Based Research on Terms of Turkish Lexicography”¹⁶ which we have previously completed. 100 terms of the 1000 lexicography terms, which we obtained as the output of our previous project, will be defined through the module created.

Key Words: corpus linguistics, definition, lexicography, module, terminology

1. Introduction

Although it may vary at what stage and for what purpose a user might consult a dictionary, the desire to learn the meaning of a lexical unit has a priority. The information category that meets this desire of

¹⁶ Eskişehir Osmangazi University Scientific Research Project - Code: 2016-019056

the dictionary user is the definition. Therefore, the answer to the "What does X mean?" question in relation to a term corresponds to definition.

The scientific language consists of terms. The formation and definition of terms are necessary to share information. Therefore, the correct definition of terms is very important for terminology studies (Girişen, 2019).

1.1. Purpose

The purpose of the project is to interactively define the terms of Turkish Lexicography which were made into lexical entries using Terminology Module of Do it Yourself Corpora Platform (DIY) for Turkish through the DIY Corpora Lexicography Module. The title of the project to be created in this sense is Interactive Term Defining Module (ITDM). This project aims to implement a module that will allow the definitions to be made interactively by multiple authors in accordance with the International Standards Institute (ISO) norms¹⁷ and the criteria of definition in terminology, especially in the new terminological dictionaries.

1.2. Scope

As the output of the previous project called "Corpus-Based Research on Terminology of Turkish Lexicography" was extracted 1000 lexicography terms. The scope of the current study is limited with 100 lexicographical terms of the CBRT-TURKLEX.

1.3. Original Value

The expected result of the current project is to develop an interactive term defining module that is accessible to multiple term definers and editors by using computer technology, and that complies with ISO standards.

Although there are defining modules in the literature, it is seen that these modules are neither interactive nor compatible with ISO standards. In addition, the terms are defined in the traditional way and they are not consistent with each other in these defining modules.

2. Method

In this study, Terminology Module (TM) that can be transformed according to linguists and researchers' research questions. The TM will tend to be accessible by multiple definers, extendable database structure, user friendly and flexible reporting system for all users. The TM is one of the modules of DIY Corpora. The website of this platform is <http://kkd.mersin.edu.tr/>.

The Turkish Lexicography Corpus (TLC) will be database of the current project. The composition of TLC includes articles, published presentations, reviews, books, doctoral dissertations, master thesis related with Turkish lexicography. The corpus contains 1003 texts that has written on Turkish lexicography. TLC comprises 42.831 sentences, 703.986 orthographic words, and 86.368 types.

¹⁷ ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application
ISO 704:2000 Terminology work – Principles and methods
ISO 704:2009 Terminology work – Principles and methods

For the terminology defined through TM:

1- Interface for labelling part-of-speech, sampling selection, defining

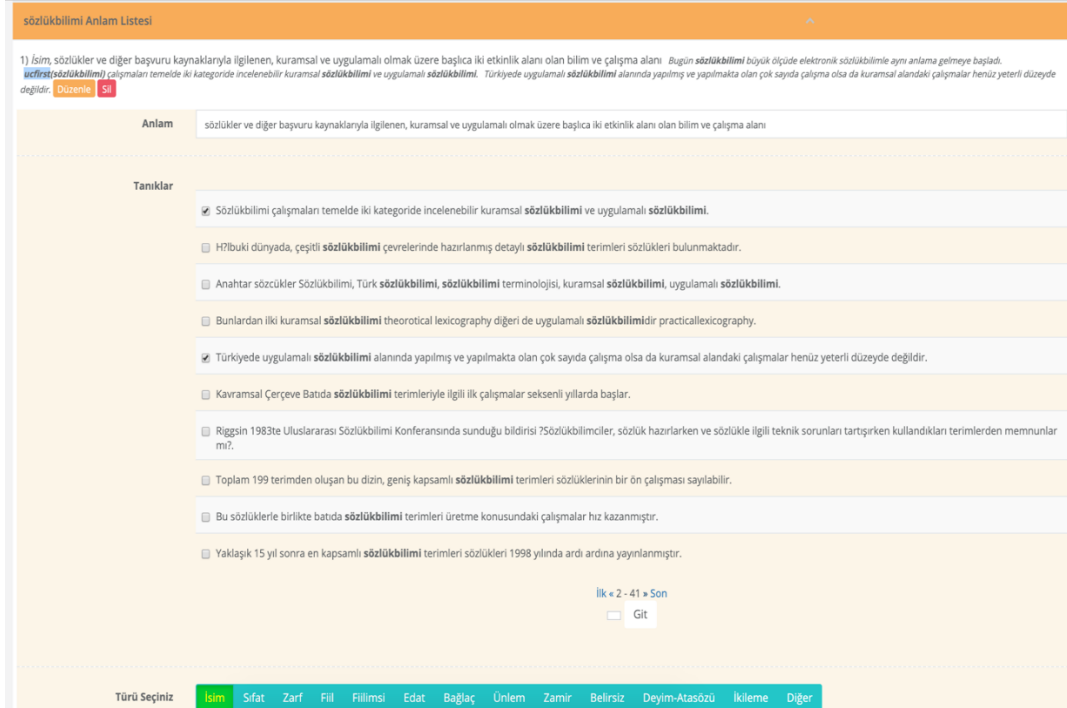


Figure1. Interface for labelling part-of-speech, sampling selection, defining

2- Entry configurations: source language, synonym/near-synonym, collocation

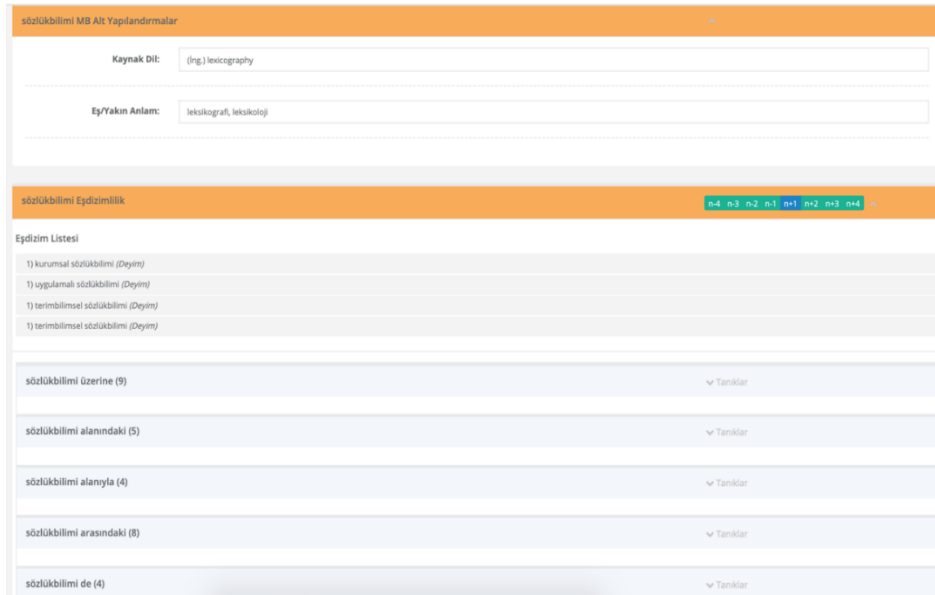


Figure.2 Entry configurations: source language, synonym/near-synonym, collocation.

3- The output of configured entries: The procedures of "lexicography" term are performed in the figure.

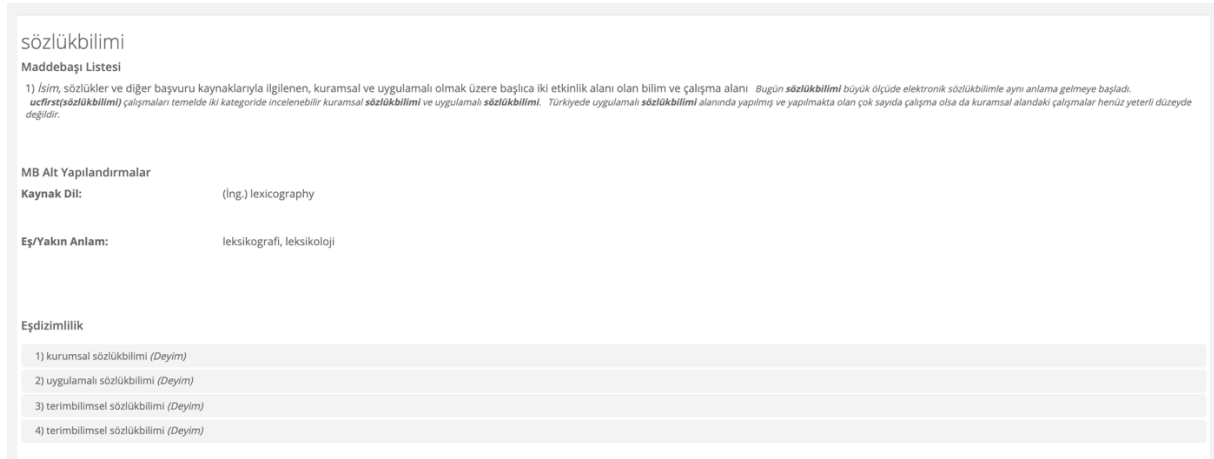


Figure 3: Configured output of an entry: the "lexicography" term

The data processing flow is as follows

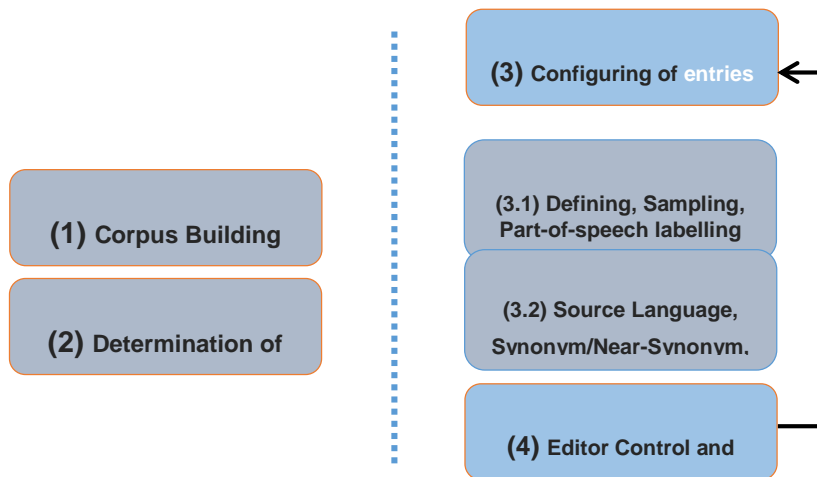


Figure. 4 Flowchart of data processing system

2.1. Definition

Definition will be performed in two stages.

First stage: The researchers will define the terms assigned to them. At this stage there will be various guiding information for definers. We can list them as follows:

i. Select the definition type

- a. Intensional definition
- b. Extensional definition
- c. Demonstrative definition
- d. Conceptual definition
- e. Lexical definition

The intensional definition will be recommended to prefer to the definers.

ii. Basic criteria in definition (intensional definition)

The definers will be asked to comply with¹⁸ the definition criteria listed below.

- a. **Preciseness:** All of the characteristics that limit the concept must be included in the intensional definition. Meanwhile, it must be noted that the definitions should not cause ambiguity, and that they should not be metaphorical, too broad or too narrow.
- b. **Conciseness:** The intensional definition should be brief but comprehensive.
- c. **Superordinate concept reference:** The intensional definition should include the superordinate i.e. general concept. When this is not possible, it may contain a more general concept.
- d. **Use of known or predefined terms:** All terms in the intensional definition should be known at large and should be defined in the source language.
- e. **Objectivity:** The intensional definition should not be subjective.
- f. **Source reliability:** The criteria such as the linguistic and technical competences of the author, the date of publication, and the publisher can provide reliability.
- g. **Target group:** The intensional definition should be able to respond to the expectations and requirements of the target group.
- h. **Scope of application:** If necessary, the scope of application should be shown.
- i. **Reference to related field:** The intensional definition should contain characteristics that reflect the perspective of the field in question.
- j. **Reference to a concept system:** The conceptual relationship of the concept defined by the other concepts of the given concept system should be expressed in intensional definition.
- k. **Correct usage of language:** Intensional definition must be in accordance with the spelling, grammar and definition writing rules of the language concerned.
- l. **Circularity/tautology:** Circularity/tautology should be avoided.
- m. **Affirmation:** It should explain what the concept is, rather than what it is not.
- n. **Avoidance of translation:** Usually it is not recommended to translate definitions or use translated definitions.
- o. **Avoidance of hidden definitions of other concepts:** Definitions of other concepts should not be included in the intensional definition.

¹⁸ The criteria were created in accordance with Cabré (1999), COTSOES (2002), Felber (1984), Pavel and Nolet (2001), Löckinger, Kockaert and Budin (2015) and ISO 704:2009.

- p. **Absence of characteristics of superordinate or subordinate concepts:** The intensional definition should not include any characteristics logically pertaining to the superordinate or subordinate concepts.

iii. Formal criteria in definition

Definitions will be required to follow the formal criteria listed below. However, this formal arrangement will be made automatically by the system.

- a. The entry term should be written in bold
- b. The entry should be written in lower case letter including the initials
- c. The entry should not be written in italic
- d. Punctuation should not be placed after the entry
- e. The definition should not be a sentence
- f. The definition should not start with a capital letter
- g. No full-stop should be placed at the end of the definition as it is not a sentence

Second stage:

The editor will check the defined terms in the control panel in this stage. The editor will evaluate the accuracy of the defined terms according to the definition criteria explained at above. The editor will give feedback to the definers about the definitions whether eligible or not.

3. Conclusion

The project team have interdisciplinary qualifications in line with the aim of the project.

The project team will include researchers from different disciplines such as linguistic researchers, lexicographers, terminologists, software experts and PhD. students. The researchers who are not a member of the project team can also contribute to the project.

In this study, the building of interactive term definition module was presented. This project is the second phase of the compiling Turkish Lexicography Dictionary initiated by “Corpus-Based Research on Terminology of Turkish Lexicography” name.

4. References

Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*. John Benjamin Publishing Company, Amsterdam/Philadelphia.

“COTSOES Recommendations for Terminology Work” (2002). *Conference of Translation Services of West European States Working Party on Terminology and Documentation*, Berne: Federal Chancellery.

Felber, H. (1984). *Terminology Manual*, Paris: UNESCO and Vienna: INFOTERM.

Girişen, N. (2019). *Terminolojide Tanım Tipolojisi*, (Basılmamış Doktora Tezi), Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

ISO 704:2009 Terminology work — Principles and methods

Löckinger, G., Kockaert H. J. ve Budin, G. (2015). “Intensional Definitions” *Handbook of Terminology*, Hendrik J. Kockaert ve Frieda Steurs (ed.). Amsterdam/Philadelphia: John Benjamin. s. 60-81.

Pavel, S. ve Nolet, D. (2001). *Handbook of Terminology*, Ottawa: Minister of Public Works and Government Services Canada.

<http://kkd.mersin.edu.tr/>

GUILLAUME BUDÉ: AN EFFICIENT TRANSFORMER OF GREEK LEXICOGRAPHY IN EARLY MODERN EUROPE

Assoc. Prof. Erman Gören, PhD
Istanbul University, Faculty of Letters
Department of Ancient Languages and Cultures
Division of Ancient Greek Language and Literature

Abstract

It is a widespread opinion held by modern scholars that Guillaume Budé (1467-1540) was one of the most efficient Greek scholars in Early Modern Europe who began to fundamentally transform Greek lexicography through his work *Commentarii Linguae Graecae* (1529). His work remains a key point in comprehending not only what the main characteristics of the transitional period of Greek lexicography from humanistic to philological approaches in his time are; but also how the traditional philological methods help us to maintain a “scientific” view as to how to select and expand a lexicographical entry. One of Budé’s greatest achievements was to induce the French King Francis I (1515-1547), to found the Collège Royal. The Collège Royal provided him with a platform similar to Ptolemy I’s *Mouseion* the institution that supplied enormous facilities to Hellenistic philologists, like Eratosthenes, Zenodotus, and the others as described by Budé himself. Budé was a new figure of *philologos* displaying both the *ethos* and *praxis* of these Hellenistic philologists. His lifestyle accorded with the new figure of *philologos* who was devoted absolutely to the Greek texts, rather than to, in Momigliano’s term, the *antiquitates* of the humanists of his epoch. Accordingly, his works, especially the *Commentarii Linguae Graecae* (1529), inspired both his pupils in Collège Royal and the great lexicographers of the next generation in France, such as Henri II Estienne (1528-1598), to work on new lexicons in a much more accurate and elegant way than exhibited by Crastone’s humanistic efforts in his *Lexicon graeco-latinum* (1476). In this paper, I will introduce this transitional and remarkable scholar, a unique figure of *philologos* in the field of lexicography, presenting an account of his life and studies. I will also indicate that the *ethos* and *praxis* of Hellenistic philologists guided Budé in constructing his identity as a *philologos* and directed in Early Modern Europe the change from the humanistic lexicography to the philological lexicography.

Key Words: Greek lexicography, Guillaume Budé, humanistic lexicography, philological lexicography, scientific lexicography

Introduction

Guillaume Budé (1468-1540) was a key figure, not only for the French, but also for many aspects of European humanism. He may be compared with the great Erasmus through his huge contributions to Greek studies, albeit because of his overloaded and often obscure style he was unable to reach as wide a public. Garandierie emphasizes his importance by indicating Budé's specific worldview which intersects definitely at a right angle to Erasmus's humanism: "Neglect of Budé is a pity ... Budé's thought is, in short, the self-consciousness of the Renaissance, the very mirror of Christian humanism" (Garandierie, 1988, p. 380). As will be discussed more thoroughly below, his view was always centralized by Christianity, but, in Garandierie's term as a "philosopher of culture," he tried to become a real transformer of ancient studies and to thereby construct a stronger Christian culture, while the consequences of his works were not limited to just the religious context. It is not possible to consider him as a Christian mystic while seeing his precisely philological works, but his aim throughout his entire life was to struggle for "climbing up some Mount Helicon" or "arising to the highest point that human spirit is able to hit."¹⁹ Budé's worldview was a specific kind of *humanismus*,²⁰ which takes the "world" as a secular object, but with a sacred aim, to fortify Christian culture through ancient literature. On the other hand, as I will argue below, Budé's influence on French humanism is not just directed by Christian ideals, rather his philological approaches initiated an explicit transformation of the entire intellectual life of Early Modern Europe.

Having graduated from Law School in University of Orléans, although his father was expecting him to become a famous lawyer, Budé "devoted himself to the pleasures of youth: riding and falconry" (Sandy, 2002a, p. 81). Yet, at the age of 23 (1491), he experienced a kind of "conversion," and this resulted in him leaving everything else and with a great ambition starting to study Greek and Latin literature. As an ardent learner, Budé "learnt Greek without a teacher" (except for a few mediocre lessons given him by Georgius Hermonymus) and often lacked suitable books (except for a few manuscripts provided by Janus Lascaris)"

¹⁹ "Heliconis seces su montis ... ad summum humani captus fastigium sublimis et tanquam volucris attollitur." *De studio literarum recte et commode instituendo* JB f. 17 (G. Budé, 1532).

²⁰ Although "it does not represent a unified view of Greek civilization; it is not the culture of the historical period which is known, with different emphases, from the work of Johann Gustav Droysen (1808–1884) and Arnold J. Toynbee (1889–1975), nor the philosophical notion which is found in the work of Matthew Arnold (1822–1888)" the term 'Hellenism' in Budé's hands, "seems to have gained, for the first time, a more generalized meaning, not linked to a specific historical culture or period, but expressive of a worldview oriented towards 'the world', rather than towards God, in the word's biblical sense and carrying negative overtones" (Lamers, 2018, pp. 205–206).

(Bietenholz & Deutscher, 1985, p. 213). In 1517, Erasmus wrote in a letter acknowledging Budé's championship of Greek studies which had been already accepted by common consent: "I now see, however, that you have been so successful that I do not believe that there is any Italian at this time who is so perverse or arrogant that he would be foolhardy enough to join arms and do battle with Budé for recognition in his field of accomplishment" (Sandy, 2018, p. 259).²¹ There was a huge difference in his competence attained between the years 1491 and 1517. How had he by then become perhaps the most efficient person in the French humanism?

Philologia had a pivotal role in the transformation of French humanism as well as of Budé himself in the transitional period from antiquarian humanism to philological humanism in the 16th century. If we accept there was a Hellenistic birth of the notion of *philologia*, we can definitely accept the same for a rebirth of it in Early Modern Europe. Then, who was the accoucheur of this rebirth? Was he Henri II Estienne, known also as Henricus Stephanus, or Joseph Justus Scaliger or Budé himself? In this paper, I will focus on the lexicographical aspect of this rebirth. The questions are how the background of Greek lexicography was, and how Greek lexicography gradually reached the level of Estienne's *Thesaurus Graecae Linguae* (1572). What was the role of Budé in this process as a lexicographer in the epoch of humanist secularization? What were the possible resources for Budé to constitute a new methodology which shed light upon the paths of the next generation lexicographers?

Method

In this study, the main source was one of Budé's own published works, especially his *Commentarii Linguae Graecae* (1529). This great lexicographical achievement was a climax of his revolutionary methodological view. Nevertheless, *De Asse et Partibus Eius Libri Quinque* (1515) was also definitely a noteworthy step that I studied in researching this intellectual ascent. Also, in his later published works, like *De Studio Literarum Recte et Commode Instituendo* (1532) and his last work, *De Transitu Hellenismi ad Christianismum* (1535), I studied Budé's unique view of Hellenism, which led him to a philological approach to lexicography. About his life, including some early biographies, the modern literature provided a fairly good picture of Budé (E. de Budé, 1884; Delaruelle, 1907; Lefranc, 1893; McNeil, 1975; Plattard, 1923). From these sources, I studied Budé's lexicographical approach comparing it with his intellectual

²¹ "Nunc adeo successisse video ut neminem hacaetate putem esse apud Italos tam improbum sibi que fidentem, qui super ista sane laude cum Budaeo congregi manusque conserere sustineat" (Erasmus, 1906, II, pp. 460).

context, including his visionary identity at the court of King Francis I and his view of Hellenism to construct a new humanistic approach for the next generation French lexicographers.

Results

It was a puzzling challenge for the French humanists to find a serviceable way to learn Greek in the last quarter of the 15th century and the first quarter of the 16th century, when Budé decided to study Greek in 1491. This big challenge could not be completely tackled for many years until Budé converted his dreams into reality at the end of 1520's. As Sandy discussed from Budé's own testimonies, even when he began to teach Greek and to translate Greek literature, a French philologist had to overcome three main obstacles in studying Greek: (1) the absence of competent teachers of Greek, (2) the absence of Greek manuscripts and printed editions, (3) the lack of reference works, such as reliable dictionaries (Sandy, 2018, p. 243). Ambitious and fond of antiquity, Budé mostly encountered the figures of *antiquarius* of the Italian humanists, a very common and almost the only form of intellectual typology in his time. As Momigliano states "... the notion of the 'antiquarius' as a lover, collector and student of ancient traditions and remains –though not a historian– is one of the most typical concepts of fifteenth- and sixteenth-century humanism" (Momigliano, 1950, p. 290).²² Hermonymus of Sparta was one of them, a typical *antiquarius*, from whom Budé took some Greek lessons. It was not possible to call him a well-educated Greek teacher, although he was a pupil of Ianos Laskaris, a Byzantine immigrant, who was one of the few people to promote Greek in Italy in the 15th century. Hermonymus was actually a copyist, for Budé, he was a supplier of Greek manuscripts rather than a real teacher. Grafton analogizes Budé's position as a student of Greek to a person who tries to learn "the German of the classical age of German literature by reading Goethe's *Faust*, part 2, with an elementary German dictionary for schoolchildren" (Grafton, 1997a, p. 153). Grafton applies the phrase "dictionary for schoolchildren" to indicate a specific dictionary which, while Budé was learning Greek, was the only available reference work. So, "until the publication of Budé's *Commentarii linguae graecae* in 1529, aspiring Hellenists in Western Europe were dependent upon the Graeco-Latin lexicon (c. 1476) of Giovanni Crastone (or Crastoni, d. after 1497), which amounted to little more than short Latin glosses of Greek words" (Sandy, 2018, p. 250).²³

Nevertheless, Budé undeniably managed to become a brilliant expert on Greek and Latin through reading voraciously the printed books and the manuscripts he could obtain access to, and by always making copious notes. Making notes was not only a process of production while he was preparing to publish his own works. As Sandy emphasizes "almost everything that he published was a work in progress, as he continued to

²² Wilamowitz indicates that the *antiquarii* were distant from practicing philology, although they served in many fields to make the Greek-Latin studies prevalent: "... what matters to us is the negative point that no concern with history or philology played any part either in the search for the old literature or in its dissemination. The humanists would long remain men of letters, publicists, teachers, but in no way did they become scholars . . . we must not expect philology of the humanists" (Wilamowitz-Moellendorff, 1921, p. 10; trans. Considine, 2008, p. 27).

²³ For a detailed view of how was Greek learning was in Western Europe from 1396 to 1529, the date of Budé's achievement in publishing *Commentarii linguae Graecae* to persuade and encourage Francis I to found Collège Royal, Botley's exhaustive study is available (Botley, 2010, pp. 1–270).

annotate his own copies of his publications, the annotations sometimes being incorporated into printed re-
editions by Budé himself and on other occasions by later editors” (Sandy, 2002a, p. 83).²⁴ Indeed, for Budé,
annotating was not only a philological practice, but also it was his deliberately imitated life style of the
Hellenistic *philologos*, in the sense of the Greek term *ethos*, as a “custom,” or a “habit.”²⁵ Budé’s ways of
annotating transformed “the printed text from a standardized product into something unique. His notes
recorded an individual’s response, laid out in a visually appealing and memorable way, to a particularly
important book” (Grafton, 1997a, p. 148). It was not a simply transformation of printed text, but a unceasing
transfiguration of Budé’s own judgments as well, which were never accepted permanently by himself.²⁶
For this reason, Budé was, as Pfeiffer states, “a φιλόλογος in the sense of Eratosthenes, who had been the
first to claim this *cognomen* for himself; it referred, according to Suetonius, to persons familiar with the
various branches of knowledge or even with the whole of the λόγος. Encyclopedic knowledge, not
eloquence, leads to true human culture; that was Budé’s conviction.” (Pfeiffer, 1976, pp. 101–102). Being
a *philologos* in the Alexandrian *Mouseion*, Eratosthenes had more than one duty to accomplish, including
teaching at the court of the Ptolemies, and editing, criticizing, and annotating texts. He in his time at the
court of Ptolemy was certainly a polyhistor, like Budé was at the court of King Francis I. Pfeiffer states that
“it would be hard to find a comprehensive term for Eratosthenes’ manifold spheres of learned activity, if
he had not coined one for himself: φιλόλογος” (Pfeiffer, 1968, p. 156), and accordingly confirms that
Budé’s position might be analogized with the original position of Eratosthenes as a *philologos*. This position
provided some advantageous conditions for the life of Eratosthenes and for other philologists in the

²⁴ Making use of Budé’s hand written marginalia Robert Estienne published a second edition of “Budé’s *Commentarii
linguae graecae*, ‘enlarged by more than a third, and corrected and improved in many places’, ... in 1548” (Armstrong,
1954, pp. 112–113).

²⁵ Budé’s philological practice and his vision to make available Greek antiquity in 16th century France, I term these
two components his *praxis*, as I will argue below, in the chapter of discussion, determined not only his ἔθος, but also
his ἦθος, namely his “disposition” or “character” (Liddell, Scott, Jones, & McKenzie, 1996, s.v.).

²⁶ In the proceeding pages Grafton demonstrates that his reading practice was a kind of child play, which has the rules
modifiable by a new knowledge from the ancient authors: “Budé, like many scholars of his day, read pen in hand,
correcting scribal and typographical blunders as he went. It was child’s play for him, for example, to see that when
Pliny discussed the permanent value of history, he suggested that it should be, not an *edema*, as the text had, but a
ktema, a “possession for always” — as the Greek historian Thucydides had stressed long before ... The rich
reconstruction of context that Budé’s form of reading required naturally endowed his encounter with each text with a
feeling of direct, close contact, gave him the feeling that he could directly imitate Roman practices in his own life”
(Grafton, 1997b, pp. 150–151).

Alexandrian court.²⁷ Budé had a similar “carefree life” both thanks to his own familial inheritance and from the facilities provided by King Francis I. But, the similar position of Eratosthenes, demanded from Budé much more strong-willed steps for him to take in his life and practices.

Budé’s life, after his “conversion” to study Greek-Latin antiquity which was maybe the strongest-willed step of his life, was completely devoted to his studies, all of his desires were oriented towards his intellectual goals. Disregarding everything except his studies, might be observed in some of his anecdotes transmitted by the biographers. For example, Eugène de Budé reports that when a servant excitedly entered his study to sound the alarm that there was a fire in the house, Budé coldly replied without even lifting his eyes from the book that he was reading, “Go and alert my wife. You know perfectly well that I do not concern myself with domestic matters” (E. de Budé, 1884, p. 22). A biography, published in 1540, the year of his death, records that Guillaume Budé “restricted himself to ‘only’ three hours of study on the day of wedding” (Sandy, 2002a, p. 81).

This considerably devoted life was an instrument for Budé to attain his humanistic goals. He was very ambitious to found a “college” to teach classical languages. “Budé thought of philology as a program for humanistic education ... This program, which reflected Budé’s desire to realize his dream of a Crown-supported institution of humanistic learning” (McNeil, 1975, p. 80). He finally realized his dream by inducing King Francis I to found the Collège Royal in 1530,²⁸ known today as the Collège de France. Budé modeled it on two trilingual institutions,²⁹ already established, one at Alcalá de Henares and the other in Leuven, while he was steadily insisting on the foundation and funding of the Collège Royal to King Francis I. In these institutions, the three languages deemed necessary for biblical study were to be represented by three separate chairs of Hebrew, Greek, and Latin. However, Budé was not only an influential member of the royal bureaucracy to convince the King to found and fund the new institution, but also, he efficiently took initiative in making the institution sustainable. As Sandy emphasizes “once Budé had mastered the language and taught it to the first two holders of the chair in Greek [Pierre Danès and Jacques Toussain, the first two *lecteurs royaux* in Greek in 1530, both of whom had been pupils of Ianos Laskaris] at the Collège de France, which he himself had been instrumental in founding, the institutionalized study of ancient Greek in France became securely established and well documented” (Sandy, 2002b, p. 78). The very first step in

²⁷ “They had a carefree life: free meals, high salaries, no taxes to pay, very pleasant surroundings, good lodgings and servants” (Pfeiffer, 1968, p. 97). For a further detailed account see Fraser, 1972, pp. 321–322.

²⁸ In February 1517, King Francis I has already declared his intention to found an institution of *lecteur royaux* devoted to study of classical language (Lefranc, 1893, p. 45). It took more than a decade to realise this intention at the court of Francis I. Finally, the king invited Erasmus to take charge of his new college. A question has been already asked even at that time: why did he not choose Budé? As Knecht ascertained correctly “the king’s choice is understandable. Budé was as good as a Greek scholar as Erasmus, possibly a more profound one, but Erasmus was the only notable scholar who satisfactorily combined the classical and Christian elements of the Renaissance and whose international reputation was commensurate with the prestige Francis hoped to gain his foundation” (Knecht, 1982, pp. 136–137).

²⁹ Both of these institutions had been established after Budé’s “conversion” to his devoted life: “The first *collegium trilinguae* appeared in 1498 within a new Spanish university at Alcalá de Henares, just east of Madrid. In 1518 the *Collegium Trilingue* [sic], inspired by Erasmus, arose in Leuven in present-day Belgium” (Turner, 2014, pp. 41–42).

setting up this “securely established and well documented” intellectual environment was publishing his lexicographical masterpiece, *Commentarii Linguae Graecae* (1529). In the preface to this work, Budé unsurprisingly analogized the new institution with Ptolemy I’s *Mouseion* (Plattard, 1923, p. 30).³⁰

Before dealing with Budé’s lexicographical masterpiece, *Commentarii Linguae Graecae* (1529) at large, it seems essential to fashion a general picture of Hellenism and to comprehend Budé’s own view of Hellenism. While Budé was taking the initiative in the founding and support for the Collège Royal, he was certainly aspiring to become a leader of French *Hellenismus*. In the beginning of the 16th century in Europe as in France, different concepts of Hellenism were quite prevalent amongst European intellectual circles, including the correspondences between the great humanists like Guillaume Budé and Desiderius Erasmus. As Henri II Estienne later followed in the steps of his magnificent predecessors, the main issue was clearly about “the compatibility of Hellenism and Christianity” (Constantinidou, 2018, p. 284). For Budé “philology leads to theology; the knowledge of divine things is the natural complement and crowning of what Budé calls ‘encyclopedia,’ and the human spirit, rising from height to height, passes effortlessly from pagan wisdom to Christian wisdom” (Delaruelle, 1907, p. 194).³¹ He developed this view of Christian Hellenism gradually, and only in his final work entitled *De Transitu Hellenismi ad Christianismum* (1535), did the formulation of this view reached to its culmination. In this work, he was moving away from the Erasmian concept of Hellenism, especially of Greek philosophy, preparing the way for Christianity, and was beginning to lay more stress upon the difference between Hellenism and Christianity. At this point, McNeil draws our attention to a significant term concerning Budé’s humanistic view: “Budé’s humanistic qualities are especially revealed in his devotion to *philologia*, which was, he remarked on several occasions, practically a second wife to him. He looked upon philology as the essence of humanism. ... In his last major work, the *De transitu* of 1535, he discusses the concept at greater length. There are two ‘philologies,’ one ‘major,’ the other ‘minor.’ *Philothoria* [which is the submit or culmination of philosophy and theology] is a higher, holier philosophy or wisdom than is mere *philologia*, which is concerned only with secular matters. Budé’s concern in 1535 was with the ‘transit’ *de minoris ad maiorem*” (McNeil, 1975, p. 78). *Philothoria* was a kind of “anagoge” which raised Christianity up to the highest point of humanity [*interdum scandens, ad summum humani*].³² Budé’s dream was to make the new *Mouseion*, namely the

³⁰ For the full French translation of the preface of *Commentarii linguae Graecae* and detail on the context see (G. Budé, 1977, pp. 346–351; Sanchi, 2006, pp. 21–37).

³¹ In his *De Asse et partibus eius libri quinque* (1515), Budé clearly displays his philological view, oriented by a Christian ideal to attain a faithful wisdom: “Animus humanus ad contemplationem sapientiae melius per cochleam iustae disciplinae scandere et intelligentius potest, quam si protinus ab infimo génère doctrinae ad summum genus discendi eompendio euaderet, scansilem disciplinarum seriem transiliens. Hoc modo Solomon encyclopediae gyro lustrasse se omnia ingeniorum monumenta significat, ut ego quidem interpretor; nosque hortari uidetur, ut per omnia philosophiae secularis et priscae dogmata, uestigia sapientiae siqua sunt (ut certe multa sunt) colligere non grauemur... Sic fiet ut cum ad studia sanctiora et monumenta sacrosancta peruenerimus, et uelut ad cubile ueritatis et sapientiae propius accesserimus, iacentem quidem illam, sed inuolutam, certius agnoscamus” (G. Budé, 1515, pp. 739–740).

³² “Eius porro montis duo vertices geminaeque minantur in coelum speculae, alteri tropologia nomen: at ea quae editior est specula, anagoge vocatur, in qua philothoria, que culmen est columenque philosophiae, interdum scandens, ad summum humani captus fastigium sublimis et tanquam volucris attollitur.” *De studio literarum recte et commode instituendo* JB f. 16v-17 (G. Budé, 1532, pp. 16–17).

Collège Royal, a watch tower [*specula*] to watch over this gradual development in the future. His Herculean work, *Commentarii linguae graecae*, would have helped to bring forth future watchmen.

The importance of *Commentarii Linguae Graecae* appears when its achievement is considered within Budé's humanistic program and when it is compared as a Greek Lexicon with its quite insufficient forerunner. Sanchi presents us with a detailed picture of *Commentarii Linguae Graecae*, in his terms, the picture of "sans doute le chef-d'œuvre d'érudition de Guillaume Budé" (Sanchi, 2003, p. 641) by comparing Budé's work with Laurentius Valla's³³ *Elegantiae Linguae Latinae* (1471). In a general view, the composition of the *Commentarii*, in fact, is based upon the Greek words without an alphabetical arrangement, with the exception of two indices, an 'Index of Greek Words and Phrases Explained in These Commentaries' and an 'Index of Latin Words and Subjects.' Each page of this continuous *Commentarii* is illustrated by the different meanings of a specific word through more or less extensive quotations from classical or late authors.³⁴ Budé had chosen to group them fairly regularly in families around the same theme or root, where nouns and verbs follow each other according to the variations of prefixes or suffixes. This apparently haphazard arrangement, the passage from one group of words to another according to variable criteria, is not only marked by titles to interrupt the flow of prose, but is blurred by parentheses or by wide digressions on various subjects, always erudite or grammatical.

Comparing some of the entries of Crastone's and Budé's works, it is possible to gain a clear impression that Budé's entries are not just much more voluminous and detailed than Crastone's, but they provide the reader with a new lexicographical point of views. For example, in Crastone, the entry of ἡ ἀγχιστεία has just three words: *affinitas propinquatio. conservatio* (Crastonus, 1476, p. a iii). In the entry³⁵ of Budé, it

³³ In Budé's own acknowledgement, Laurentius Valla was a significant scholar: "Ego uero Laurentium Vallensem egregii spiritus uirum existimo seculi [*sic*] sui imperitia offensum, primum Latine loquendi consuetudinem constituere summa religione instituisse: deinde iudicii acrimonia singulari, euro profectus quoque diligentiam aequasset, in eam superstitionem sensim delapsus esse, ut & sese ipse & alios captiosis obseruationibus scribendique legibus obligaret" (Sanchi, 2003, pp. 649–650).

³⁴ Budé was moving away from his Italian and Byzantine antecedents as a lexicographer via "defining Greek words by taking examples from classical texts rather than from the ancient and Byzantine collections of rare words employed by Italians and their Byzantine teachers with scarcely a reference to a classical Greek author" (Sandy, 2002b, p. 60).

³⁵ "Est enim Ἀγχιστεία ius successionis legitimae. Isaeus [*De Hagnia* 2.3-4], Ὁ νόμος δίδωσιν ἀγχιστείαν τοῖς πατρὸς ἀνεψιοῖς, Lex refert haereditatem aminitis, vel patruelibus, vel agnatis. varie enim accipitur ἀνεψιός. Et Ἀγχιστεύειν. Idem [11.2-3], Γνώσεσθε τοῦθ', ὅτι ἐμοὶ μὲν ἀγχιστεύειν, τοῖς δ' ἐξ ἐκείνων γεγονόσιν οὐκ ἦν, id est, iure propinquitatis haereditatem cernere aut petere. Et Ἀγχιστεῖς οἱ συγγενεῖς, agnati et cognati. Ἀγχιστεία etiam est

was not a single entry, but a whole series of entries. He juxtaposes sequentially the words of the same root, even those words in the same semantic field. Budé according to his concept of Hellenism, tries to display the relationship of juridical and religious languages. After *ankhisteia* he discusses the verb *ankhisteuein* in the same root, as the act of the concept of *ankhisteia*, then *ankistheus* succeeds as the agent of the first one. But Budé does not finish his comment, but continues with the related words, like *ankhitheos* in the context of religious terminology. *Ankhōmalos* with two excerpts from Plutarch and Thucydides is followed by two words in the same semantic field: *[kata] genos* and *[kata] diathēkas*. Indeed, *prima facie*, Budé's arrangement might look like constructed by arbitrary choices. Yet, Sandy gives another and similar example³⁶ to demonstrate that “the difference between Crastone's glossary and Budé's ‘commentaries’ is immense” (Sandy, 2002b, p. 67). How can this “immense difference” be defined? Sanchi thinks that “the apparent disorder that characterizes the organization of the text of the *Commentaries* reveals the great complexity of Budé's linguistic research and the novelty of his conceptions. The best Italian erudite tradition inaugurated by Valla, taken over by Poliziano, was renewed by Budé” (Sanchi, 2003, p. 653).

ius proxime ad aliquem ac cedendi. Plut. de Flamine [Plutarch, *Numa* 8] dixit, τούτω περι τὸ θεῖον ἀγχιστεῖαν. Et Ἀγχιθεος cui in primis sacra facere ius est, oraculaque petere. Λουκ. [Lucian, *De Syria Dea* 31.6-8] Οὐ μέντοι πάντες εἰσέρχονται ἱερέες, ἀλλὰ οἱ μάλιστα ἀγχιθεοὶ τε εἰσὶ, καὶ οἷσι πᾶσα εἰς τὸ θεῖον μέλεται θεραπεία. Ἀγχώμαλος ὁ παραμικρὸν ἴσος, ferme aequalis, παρὰ τὸ ἀγχι καὶ ὀμαλός. Plut. Caes. [Plutarch, *Caesar* 42.2] ἦν δὲ καὶ τὸ τῶν πεζῶν πλῆθος οὐκ ἀγχώμαλον, ἀλλὰ τετρακισμύριοι καὶ πεντακισχίλιοι παρετάπτοντο δισμυρίοις. Thucy. [Thucydides 3.49.1-2] καὶ ἐγένοντο ἐν τῇ χειροτονίᾳ ἀγχώμαλοι, id est, Pene parem utraque sententia numerum tulit. κατὰ γένος etiam dicitur: sicut κατὰ διαθήκας, ex tabulis petere. Demosth. [Demosthenes, *Contra Macartatum* 5] καὶ τοῦ κήρυκος κηρύττοντος, εἴ τις ἀμφισβητεῖν ἢ παρακαταβάλλειν βούλεται τοῦ κλήρου τοῦ Ἁγνίου ἢ κατὰ γένος ἢ κατὰ διαθήκας, οὐκ ἐτόλμησεν παρακαταβαλεῖν, Cum praeco pronunciasset licere qui vellet de haereditate ambigere Hagniae, aut propinquitate, aut ex testamento. Vel, licere si cui videretur bonorum possessionem agnoscere, aut unde agnati, aut secundum tabulas: haereditatemque petentem, sponsione aut sacramen to agere, hic sponsione facere non ausus est” (G. Budé, 1530, pp. 107.33-108.20).

³⁶ Sandy gives an example that “Crastone's glossary four isolated Latin words [*sacrificium. Sacrum. purification, sanctimonia*] to serve to define ἡ ἀγιστεῖα. Budé devotes some thirty entries to matters relating to sacrifice. For each entry he cites several sources; each source is quoted so that the word in question can be viewed in its syntactic context; he notes syntactic peculiarities such as the use of the dative or genitive with a verb or noun, and figurative uses of words such as ‘the sacred rites of Love’ in Plato's *Symposium*; at times Budé even corrects the transmitted text, e.g. ‘where perhaps ... is to be read instead of ...’” (Sandy, 2002b, p. 67).

However, if there actually was, there must have been in Budé's mind, as Sanchi indicates, that which was the "certaine organisation interne" (Sanchi, 2003, p. 645) of the *Commentarii Linguae Graecae*.

Budé had relatively better lexicographical examples published in Latin than in Greek, and Valla's work was not the only one. Like Valla's *Elegantia*, as Ann Moss indicated, Perotti's *Cornu copiae* has no alphabetical order, and Budé's *Commentarii Linguae Graecae* was a similar effort: "A total culture is brought to life, explained, and grasped in all its complex detail, and its language, its concepts, and its material objects are authenticated from textual evidence. Perotti's method of exploring verbal proximities conveys the sense of being inside a language, of following its natural associations ... Alphabetization may be convenient, but it is a search tool to be used from outside a system. Humanists like Perotti want their readers to be fully assimilated within the system" (Moss, 2003, pp. 20–21). This system that prior to the alphabetic orders of prospective lexicons of the late 16th century, like Henry Estienne's *Thesaurus Graecae Linguae* (1572), was created for addressing its reader as a person who should read within the culture, not just browsing a single word. Budé embodied his philological erudition inside this kind of system to promote his humanistic ideals.

This kind of "system" may be analogous to the efforts to constitute a systematic classification in another desired field of study which had a philological starting point for a new period in the last quarter of the 15th century in the European renaissance, and also nurtured abundantly from Greek antiquity: the studies on botanical plants. In Europe, the earliest botanical "gardens had been founded in Pisa, Padua and Florence by 1546, and the next twenty years saw their establishment in Ferrara, Sassari, Bologna and many other places in Italy" (Morton, 1981, p. 121). Yet, surprisingly the first diligence to compile a list of plant inspired by Theophrastus' and Pliny's botanical taxonomies appeared in France. While Budé was studying on his *Commentarii Linguae Graecae*, another humanist, Jean Ruel (1474-1537), a physician fluent in Greek and Latin, was studying how to construct a botanical systematic at the court of King Francis I. Ruel published a Latin translation of Dioscorides' *De Materia Medica* in 1516. In his *De Natura Stirpium* (1536), the cost of its first edition had been paid by Francis I, Ruel "attempted to give a systematic descriptive morphology of plants for which he lists a fairly extensive terminology: this gives a valuable indication of the technical vocabulary of contemporary botanists, pre-dating the glossary of terms published by Fuchs" (Morton, 1981, p. 122). Ruel's work had an intention to transmit the botanical taxonomies of Theophrastus and Pliny by describing about 600 plants, but his effort was not sufficient to construct a scientific "system" but it enabled the first systematic fruits in Leonhart Fuchs' *De Historia Stirpium Commentarii Insignes* (1542). In the strict sense the "scientific" classification of botanical plants waited until the 18th century with the publication of Carl Linnaeus' (1707-1778) *Species plantarum* (1753), which presented a hierarchical

classification of plant species. His system for naming, ranking, and classifying organisms is still, with many changes, in wide use today.

As a profound polyhistor of the renaissance, Budé was also interested in botanical vocabulary, in his “second notebook, entitled on the binding of the volume *Herbae, Frutices et Pigmenta*, there are many examples of botanical vocabulary of daily use, where French rubs Latin, and often Greek” (Sanchi, 2006, p. 117). Budé and Ruel were certainly aware of each other’s works. Ruel had also few other instruments than the ancient authors to compile his *De Historia Stirpium*, as Budé employed in his *Commentarii Linguae Graecae*. Morton commented on Ruel that he “did not apply the new terminology consistently and failed to relate it to descriptions taken from ancient writers, which are quoted alongside without discrimination or comment. In this he was no doubt simply reflecting the state of botanical thought at the time: he was a compiler and transmitter, not an original thinker” (Morton, 1981, p. 122). The importance of Budé’s contribution through his *Commentarii Linguae Graecae* can be clarified by comparing him not only with contemporary experts on Greek studies, but also with contemporary scholars in all fields.

Discussion

Budé’s works were milestones for the transitional period from antiquarian humanism to the philological approaches of the late 16th century. Considine emphasizes this remarkable epoch, in which Budé strongly influenced his pupils, stating “the philological study of ancient texts, and the philological lexicography that went with it, really began in the first half of the sixteenth century” (Considine, 2008, p. 26). The vast distinction between Crastone’s glossary and Budé’s ‘commentaries’ reveals that Budé’s devotion to *philologia* not only converted him from a hunting vagabond in the wild into a precise scholar of the ancient civilization, but also irreversibly transformed and led French humanism into an accurate and refined philological proclivity. It is of common consent that “Budé transformed Greek lexicography. There is scarcely a page in Henri II Estienne’s Graeco-Latin dictionary that does not contain a reference to Budé” (Sandy, 2018, p. 250). The building of this transformation had two main buttresses: (1) Budé’s devotion to *philologia* as a new figure of *philologos* like the Hellenistic Eratosthenes [*ethos* or *ēthos*], (2) Budé’s initiative as a teacher and/or visionary to promote his humanistic view [*praxis*].

Budé’s devotion to *philologia* was not just a professional ambition, but the anecdotes reported by the biographers show he was of the “habits” or “customs” [*ethos*] of reading and annotating without allowing any interruption from daily life of his studies. These “habits” and “customs” thereby became his

“disposition” or “character.” Namely, Budé like his Hellenistic predecessor Eratosthenes, constructed a life-style for himself, containing both *ethos* and *ēthos*, out of *philologia*.

He was not content with this life style by just being engaged with his own publications, he showed the results of this life style in concrete actions, cultivating many pupils and inducing the King to found a productive institution, the Collège Royal, known today as the Collège de France. Hence, his practice [*praxis*] was not only and solely a scholarly endeavor, but was integrated, furthering the aim of promoting a new philological humanistic view. Namely, Budé made his scholarly practice and his visionary deeds, out of *philologia*.

So, why did not Budé apply to his *Commentarii Linguae Graecae* an alphabetic order? Budé never claimed that his *Commentarii* was a lexicon. It seems it was a deliberate and considered choice by Budé not to apply to his “commentaries” an alphabetic order. The aim of the “commentaries” was actually keeping the audience within the spirit of the ancient culture, not just draw their attention through a single lexicographical entry. But this conscious decision gave way to a new generation of philologists, and in less than 50 years, Henri Estienne’s *Thesaurus Graecae Linguae* (1572) satisfied the need for a lexicon with an alphabetical order. As Stevens rightly indicates “between Guillaume Budé and Henri Estienne, the two pylons of the Renaissance, lies the most brilliant period of French Hellenism. Just as the former represents the learning of Lascaris, Aleandre, Tissard, and Tousain, the latter, prodigious son of an erudite father, is the finished scholar who profits by the labors of his predecessors” (Stevens, 1950, p. 241).

Conclusion

Valla’s and Perotti’s works were imitable and improvable examples for Budé, but does this conclusion provide us with an answer to the question of what the “certain organisation interne” of the *Commentarii Linguae Graecae* is. When we are searching for what was the genuine aim of his ostensibly capricious attitude, while he was not applying any alphabetical order in his *Commentarii Linguae Graecae*, we solely encounter a steadfastly annotating *philologos*. I conclude that Budé’s “certain internal organization” in his *Commentarii Linguae Graecae* was analogous to the contemporary though concerning botany. Early botanical gardens were organized according to the relationship of the plants to each other. Symbiosis was the primary criterion to arrange and juxtapose the plants and their living environments. I think Budé’s reliance on not applying an alphabetic order was definitely a deliberate choice, but this was also a necessity. Before presenting the living environments of the vocabulary, it was impossible to construct a lexicographical taxonomy. Budé prepared the path for the first complete lexicographical Greek taxonomy,

embodied in Henri II Estienne's *Thesaurus*. Budé was an extraordinary talent, but he was a brilliant star of his time.

How this “certain internal organization” can be analyzed thoroughly, how can the lexicographical strategies in each entry be detected? Is it possible to comprehend the relationship between the strategies and the pivotal themes which Budé applied in his un-headed ‘Commentaries’? What are the details of the interaction between the taxonomical thought in different fields, especially in the 16th century? These questions cannot be pursued here, but they deserve a more serious treatment elsewhere. Indeed, a good deal of work still needs to be done to smooth out the imbalances and fill the lacunae in our knowledge of the subject. Budé's phenomenal philological works captivated and still captivate many scholars. Although his readers in his time were enchanted, and modern readers are still struck with admiration for his overloaded and often obscure style, the superiority of the *Commentarii Linguae Graecae* over former lexicographical efforts was evident to everyone, and without doubt remains so.

References

- Armstrong, E. (1954). *Robert Estienne, Royal Printer: An Historical Study of the Elder Stephanus*. Cambridge: Cambridge University Press.
- Bietenholz, P. G., & Deutscher, T. B. (1985). *Contemporaries of Erasmus: A Biographical Register of the Renaissance and Reformation. Volume 1, A-E*. Toronto/Buffalo: University of Toronto Press.
- Botley, P. (2010). Learning Greek in Western Europe, 1396-1529: Grammars, Lexica, and Classroom Texts. *Transactions of the American Philosophical Society*, Vol. 100, pp. 1–270. <https://doi.org/10.2307/41062660>
- Budé, E. de. (1884). *Vie de Guillaume Budé, fondateur du Collège de France (1467-1540)*. Paris: Émile Perrin, Libraire Éditeur.
- Budé, G. (1515). *De asse et partibus eius lib[r]i quinq[ue]*. [Paris]: Vϛundantur in ϛdibus Ascensianis.
- Budé, G. (1530). *Commentarii linguae Graecae*. Basileae: In aedibus [Johann Bebel].
- Budé, G. (1532). *De studio literarum recte et commode instituendo: ad invictissimum et potentissimum principem Franciscum Regem Franciae*. Parisiis: Excud. J. Badius Ascensius.
- Budé, G. (1977). *Correspondance, Tome I: Les Lettres grecques—adjectis paucis e latinis* (G. Lavoie & R.

Galibois, Eds.). Sherbrooke: Centre d'Études de la Renaissance, Université de Sherbrooke.

Considine, J. P. (2008). *Dictionaries in Early Modern Europe: Lexicography and the Making of Heritage*. Cambridge/New York: Cambridge University Press.

Constantinidou, N. (2018). Constructions of Hellenism Through Printing and Editorial Choices: The Case of Adrien de Turnèbe, Royal Lecturer and Printer in Greek (1512–1565). *International Journal of the Classical Tradition*, 25(3), 262–284. <https://doi.org/10.1007/s12138-018-0470-1>

Crastonus, J. (1476). *Lexicon graeco–latinum*. Milan: Bonus Accursius.

Delaruelle, L. (1907). *Guillaume Budé. les origines, les débuts, les idées maîtresses*. Paris: Librairie Honoré Champion.

Erasmus, D. (1906). *Opus epistolarum Des. Erasmi Roterodami* (P. S. Allen, H. M. A. Allen, H. W. Garrod, & B. Flower, Eds.). Oxonii: in typographeo Clarendoniano.

Fraser, P. M. (1972). *Ptolemaic Alexandria. Volume 1: Text*. Oxford/New York: Clarendon Press.

Garanderie, M.-M. de la. (1988). Guillaume Budé, A Philosopher of Culture. *Sixteenth Century Journal*, 19(3), 379–388. <https://doi.org/10.2307/2540469>

Grafton, A. (1997a). Is the History of Reading a Marginal Enterprise? Guillaume Budé and His Books. *The Papers of the Bibliographical Society of America*, 91(2), 139–157. <https://doi.org/10.1086/pbsa.91.2.24304538>

Grafton, A. (1997b). Is the History of Reading a Marginal Enterprise? Guillaume Budé and His Books. *The Papers of the Bibliographical Society of America*, 91, 139–157. <https://doi.org/10.2307/24304538>

Knecht, R. J. (1982). *Francis I*. Cambridge: Cambridge University Press.

Lamers, H. (2018). Constructing Hellenism: Studies on the History of Greek Learning in Early Modern Europe. *International Journal of the Classical Tradition*, 25(3), 201–215. <https://doi.org/10.1007/s12138-018-0467-9>

Lefranc, A. (1893). *Histoire du Collège de France, Depuis des Origines Jusqu'à la Fin du Premier Empire*. Paris: Librairie Hachette.

Liddell, H. G., Scott, R., Jones, H. S., & McKenzie, R. (1996). *A Greek-English Lexicon*. Oxford/New York: Clarendon Press.

McNeil, D. (1975). *Guillaume Budé and Humanism in the Reign of Francis I*. Genève: Droz.

Momigliano, A. (1950). Ancient History and the Antiquarian. *Journal of the Warburg and Courtauld Institutes*, 13(3/4), 285–315. <https://doi.org/10.2307/750215>

Morton, A. G. (1981). *History of Botanical Science: An Account of the Development of Botany from Ancient Times to the Present Day*. London/New York: Academic Press.

Moss, A. (2003). *Renaissance Truth and the Latin Language Turn*. Oxford/New York: Oxford University Press.

Pfeiffer, R. (1968). *History of Classical Scholarship: From the Beginning to the End of the Hellenistic Age*. Oxford: Clarendon Press.

Pfeiffer, R. (1976). *History of Classical Scholarship: 1300 to 1850*. Oxford: Clarendon Press.

Plattard, J. (1923). *Guillaume Budé (1468-1540) et les origines de l'humanisme français*. Paris: Les belles lettres.

Sanchi, L.-A. (2003). Guillaume Budé et ses Devanciers Italiens: A Propos des Commentaires de la Langue Grecque. *Bibliothèque d'Humanisme et Renaissance*, 65(3), 641–653. <https://doi.org/10.2307/20680654>

Sanchi, L.-A. (2006). *Les commentaires de la langue grecque de Guillaume Budé : l'œuvre, ses sources, sa préparation*. Genève: Droz.

Sandy, G. (2002a). Guillaume Budé: Philologist and Polymath. A Preliminary Study. In *The Classical Heritage in France* (pp. 79–108). Leiden/Boston/Köln: Brill.

Sandy, G. (2002b). Resources for the Study of Ancient Greek in France. In G. Sandy (Ed.), *The Classical Heritage in France* (pp. 47–78). Leiden: Brill.

Sandy, G. (2018). Guillaume Budé and the Uses of Greek. *International Journal of the Classical Tradition*, 25(3), 241–261. <https://doi.org/10.1007/s12138-018-0469-7>

Stevens, L. C. (1950). How the French Humanists of the Renaissance Learned Greek. *PMLA*, 65(2), 240–

248. <https://doi.org/10.2307/459467>

Turner, J. (2014). *Philology: The Forgotten Origins of the Modern Humanities*. Princeton: Princeton University Press.

Wilamowitz-Moellendorff, U. von. (1921). *Geschichte der Philologie*. Leipzig: B.G. Teubner.

THE USER IS KING: ADVICE TO LEXICOGRAPHERS OF LEARNER'S DICTIONARIES³⁷

Donna M.T.Cr. Farina

New Jersey City University

Marjeta Vrbinc

University of Ljubljana

Alenka Vrbinc

University of Ljubljana

Abstract

This paper presents a study that was carried out through a collaborative project between Slovenia and the USA. The findings are based on interviews with economics students, advanced English learners, from the University of Ljubljana. Subjects were first asked general questions about their habits of dictionary use; the researchers then tested their look-up ability as well as their perceptions of the utility and quality of definitions and illustrative examples. Nine short sample contexts were used for the tests; these contexts contained a clearly-marked common word in an infrequent sense, such as *sharp* in the meaning ‘fashionable’. Subjects were asked to read each context and then indicate whether they knew the meaning of the target word. Next, they had to locate the appropriate sense in the online edition of *Merriam-Webster Learner’s Dictionary* (MWLD) and contrast their initial, ideas about the meaning in context to the definition(s) they found. A think-aloud format enabled the researchers to follow the students’ look-up process and note their problems. The students expressed a broad spectrum of opinions on the information provided in the dictionary as well as on the dictionary’s methods of information presentation. They suggested ways the dictionary could be more helpful to users, and their comments form the basis for recommendations in the following areas: the improvement of drop-down menus; the inclusion of numerous forms of the target word in search tools (such as drop-down menus) as well as in illustrative examples; the optimal number and length of illustrative examples; the inclusion of information in square brackets within illustrative examples; the use of italics, boldface, and colors in the online environment; and the immediate or optional accessibility of different types of information.

³⁷ An expanded version of this article is forthcoming in *International Journal of Lexicography*, 32 (2019) under the title: “Problems in Online Dictionary Usage for Advanced Slovenian Learners of English.”

Key Words: online learner's dictionary, habits of dictionary use, look-up abilities, dictionary definitions, illustrative examples.

Introduction

A qualitative study carried out in March 2018 examined the habits, impressions, and look-up challenges of online learner's dictionary users. This study was a joint project between Slovenia and the United States (see Acknowledgements). The nine subjects were students at the University of Ljubljana, advanced learners of English in the Faculty of Economics, majoring in different areas of business and economics. While these students used English on a daily basis, they are by no means specialists in the English language.

The study focused on many aspects of the dictionary look-up process: the multifarious actions taken during a search for nine infrequently-used word senses in the American *Merriam–Webster Learner's Dictionary* (MWLD), an online dictionary. Here we provide information about the expectations these users brought to the table as they approached this dictionary. We also present recommendations for the improvement of diverse aspects of online learner's dictionaries, based on the perceptions of the nine study subjects.

Method

English capability of participants

Generally, before beginning their university studies, students in Slovenia reach the level of B2 in the Common European Framework of Reference for Languages (CEFR) (*Učni načrt Angleščina*, 2008; Bitenc Peharc and Tratnik, 2014; Ilc and Stopar, 2015). This is approximately Advanced Mid in the American Council on the Teaching of Foreign Languages (ACTFL) proficiency scale (*Assigning CEFR Ratings to ACTFL Assessments*, n.d.). All nine volunteers for this study were in their third or final year of undergraduate studies in the Faculty of Economics, in a program that requires them to use English on a daily basis. It is safe to assume that they were all B2 or higher, advanced in English and fairly homogeneous in their English skill level. In fact, the researchers' perception was that the subjects were at this level or higher.

Target dictionary

The online *Merriam–Webster Learner's Dictionary* (MWLD) was chosen because it is not known by university students in Slovenia; in fact, no participant had ever used it or was familiar with it. Slovenian students generally use learner's dictionaries produced in Great Britain, so we wished to test them on a work emphasizing American English.

Target words/senses; target contexts

Initially, we chose numerous common words used in standard American English in infrequent senses. Then, our word list was shortened based on the availability of suitable contexts (i.e., texts where each word/infrequent sense is used) until we arrived at the nine words/senses selected for this study. We sought contexts written in a relatively uniform style, not in academic English but from texts that most educated advanced learners would grasp. The contexts reflect contemporary language in the U.S.; the time range of the nine contexts spans the years 2000 to 2018. Seven of these contexts—for three nouns, three verbs, and

one adjective—are from *The New York Times* newspaper. Two contexts, for the adjectives *mean* and *rich*, are from an American non-fiction text and a fictional text, respectively.

The study purposely included senses that would challenge the subjects' dictionary look-up skills. While the target words are common and frequent ones that these students would be expected to know, the subjects were required to locate an infrequent sense of each word in the dictionary. In fact, each participant was familiar with the most common meaning(s) of each target word. They usually did not know or had a vague comprehension of the infrequent sense in this study's focus. What is more, even when the students claimed to be familiar with an infrequent target sense, their proposed definitions (prior to dictionary look-up) were for the most part wrong.

Design of the study; themes; interview questions

This qualitative study with anonymous participants was intended to elicit detailed information about dictionary use; it used semi-structured interviews along with the direct observation of dictionary look-up tasks (Hatherall, 1984; Qu and Dumay, 2011). There were fourteen scripted interview questions related to pre-identified broad themes:

Broad Theme 1: Habits of Dictionary Use

Broad Theme 2: Look-up Ability of Participants

Broad Theme 3: Perceptions of Utility and Quality of Definitions

Broad Theme 4: Perceptions of Utility and Quality of Illustrative Examples

During the interviews, the scripted questions were supplemented with “probes designed to elicit more elaborate responses” (Qu and Dumay, 2011, p. 246). In addition, think-aloud protocols were used (Wingate, 2002), where the subjects were asked to verbalize what they were doing while looking up the nine words/senses online in the MWLD.

The first six interview questions were general; they were asked before a subject was given the nine contexts to read. These quasi-“introducing questions” (Qu and Dumay, 2011, pp. 249–250) helped to provide background for what would transpire later during the interview (Kvale, 1996). Three of the questions covered dictionary usage habits (preferences and frequency of use) and two inquired about subjects' satisfaction with what they usually find in dictionaries and how quickly they find it; one question asked what they dislike or miss in the dictionaries they use (Broad Theme 1: Habits of Dictionary Use).

After answering the six general questions, each participant read a target context, was asked whether they knew the meaning of the word used, and whether they could tell us the meaning. Then each subject proceeded to look up the word. Students searched online in the MWLD for the correct sense to correspond to the use in the target context (Broad Theme 2: Look-up Ability of Participants). Next, the students were asked whether their initial definition was correct, and how their initial idea or guess about the word's meaning compared to what the dictionary had. We then asked three questions related to the illustrative examples: their usefulness, what the subject liked about them, and how they could be made better; another question, about information in square brackets “[]”, asked whether the example could be understood without this information (Broad Theme 4: Perceptions of Utility and Quality of Illustrative Examples). Finally, participants were asked to indicate any part of the dictionary information that was the most useful to them in understanding the meaning of their context (Broad Theme 3: Perceptions of Utility and Quality

of Definitions). The researchers asked follow-up questions about subjects' experience during the look-up task, about the MWLD, and other areas (all four Broad Themes).

Results

Broad theme 1: Habits of dictionary use

Most subjects use a dictionary once per week or more often; however, one participant uses dictionaries rarely (once per month) and relies mostly on Google. Those who use dictionaries choose whatever online dictionary comes up first on a Google word search, in line with the findings of Lorentzen and Theilgaard (2012). They use to a much lesser extent non-learner's English monolingual dictionaries that they know by name. None of the students had ever used the target dictionary for this study, the *Merriam–Webster Learner's Dictionary* (MWLD); they had never heard of it.

Eight of the nine subjects were satisfied with the monolingual dictionaries that they use. The sole participant who expressed dissatisfaction was not discussing dictionaries but Google Translate. All but one student stated that they find what they are looking for quickly. The answers to questions about what subjects miss or dislike in dictionaries yielded highly individualized answers. One student talked about too much complexity and too many abbreviations in dictionaries; another complained that a certain dictionary provides too much information and is poorly organized; another mentioned unclear definitions; another said that there are too few examples of word use. One participant stated that s/he had never thought about our questions since s/he rarely uses a dictionary and uses Google Translate.

Broad theme 2: Look-up ability of participants

The discussion of habits above indicates that the study subjects do use dictionaries, although not as often as lexicographers might like. The subjects' look-up skills were often not strong enough for them to benefit fully from interaction with a dictionary. While it has been argued that the digital format of a dictionary provides easier access to information (Lew and de Schryver 2014), these nine users often did not demonstrate a deep understanding of the information they accessed. This study identified significant problems that can be attributed to students' look-up skills. While these problems are sometimes interrelated and two or more of them can contribute to a difficult or failed look-up experience, any single problem can cause a user to take missteps that cost time and/or lead to incorrect sense identification.

A good example of a skills-based problem is attention to part of speech. Our subjects often did not notice the part of speech of the words in their contexts; then, when they began looking up words, they continued not to notice part of speech. The subjects could tell the difference between, for example, a noun and a verb, but they were not attentive to that difference and often ignored it. When the subjects read a word in a context, they were interested solely in deciphering the meaning:

Context 5, *plug* (noun) (excerpt)

[...] But you can't shake the feeling that it's all just a big **plug** for Microsoft's music store.

[...]

Another skills-based problem had to do with the form of the target word. It goes without saying that words in real contexts, particularly verbs, do not always appear in the dictionary in their canonical form; the form

of the lemma often does not match the form of the word in context. This presented a huge problem for our subjects and often was a factor (or *the* factor) leading to their failure to find the correct target sense in the dictionary (see also Lew and de Schryver (2014)). All three of the verb forms in this study—*taxed*, *fixing*, and *scoring*—presented look-up problems for the users and the form of the target verb was the main driver in subjects’ efforts to locate the correct meaning. Seven of the nine participants did eventually find the correct sense of the verbs as used in their context, but it was time-consuming and difficult for them.

Broad Theme 3: Perceptions of utility and quality of definitions

One question asked participants to indicate any part of the dictionary information that was the most useful to them in understanding the meaning in the context. This question allowed us to understand subjects’ perceptions of definitions. In addition, we received some information about subjects’ views on definitions from five other scripted questions, from our own observations, and from unscripted questions.

For the verb *score*, eight students identified the correct sense (“slang : to buy or get (illegal drugs),” and five of these mentioned the usefulness of the phrase *illegal drugs* appearing in parentheses within the definition. The context was as follows:

Context 3 *score* (verb) (excerpt)

Back when Patrick had a job at an auto-parts store and as a banquet server, his morning routine involved driving to Lawrence before work and **scoring** his daily fix. [...]

For the noun *plug*, four of the eight students who found the correct sense mentioned the grammatical collocation *often + for* as useful information within the MWLD definition; it is probable that this helped them understand all three illustrative examples which contained *a plug for*:

- I heard a *plug for* that café on the radio.
- He **gave a plug for** [=talked about] his new film during the interview.
- She **put in a plug for** the band's new album on her radio program.

Three of the eight subjects who were correct in their choice of sense for *plug* also mentioned the illustrative examples as useful, along with the definition.

The perception of definition utility does not always mean that the correct sense was selected by the subjects; sometimes students who chose the *wrong* sense nevertheless had a favorable view of a definition’s utility. Participants who picked an incorrect sense provided indirect evidence that the definition might not have been very useful. However, there are many reasons for choosing an incorrect sense that are not due to faulty definitions. During our informal conversations with students, they added comments reinforcing the conclusion that overall they were satisfied with the defining style of the MWLD, as well as with the dictionary taken as a whole.

Broad theme 4: Perceptions of utility and quality of illustrative examples

Participants responded to several questions about examples: what they liked about them, whether they illustrated the word’s meaning effectively, how they could be improved, whether material in square brackets within the examples was useful, and whether they would have understood an example without the material

in square brackets. There were six aspects of the illustrative examples that the students noticed and commented on frequently; these are discussed below.

Length of illustrative examples. The participants stated most often that the examples in the dictionary were too short. In particular, they felt that examples were too short when they were not in full sentences. For example, for the noun *pitch*, four students said that *an advertising pitch* is too short as an example. One student stated that s/he doesn't like "short segments;" another said that the example "lacks context." Another student considered the example too general: "You can advertise a lot of things;" "If you don't know what *pitch* is, it's not helpful." In only a single instance during the full study did a student say that an example was too long. Likewise, in a single instance one student claimed that s/he liked short examples, but never pointed to a specific example that s/he said was too long.

Number of illustrative examples. The students in this study frequently expressed the desire for a greater number of examples than were provided. For six out of the nine senses studied, one or more students stated that they would have preferred more examples. We observed that when the subjects said they wanted a greater number of examples, they often meant that the examples provided did not, in their opinion, cover their specific context.

Similarity of wording in illustrative examples to wording in contexts. We observed that the most difficult task for these participants was to connect the infrequent sense of their context to the related information in the dictionary. Because this could be so challenging, they mentioned often the similarity or the dissimilarity of their contexts to the examples of the MWLD. In many cases where the researchers considered an illustrative example to be obviously similar to a given context, the participants thought the example was dissimilar. Subjects made comments related to similarity/dissimilarity for six of nine target words/senses. The researchers were surprised by the degree of sensitivity that the subjects had to very small and (for us) irrelevant differences between the contexts and the dictionary examples.

Information in square brackets within illustrative examples. The prevailing view among our participants was that they preferred having information in square brackets within the examples. For example, for the verb *tax*, one student said that s/he would have understood the meaning of the first two examples without the additional information in brackets. Nevertheless: "Additional explanations are helpful because the words are explained once again in a simple way." For the fourth example, the brackets were essential for this subject's understanding:

- That job really *taxed* our strength. [=required us to use a lot of physical effort]
- All this waiting is *taxing* my patience. [=is making me lose my patience]
- puzzles that *tax* your brain
- You can have an enjoyable vacation without *taxing* your budget. [=without having to spend a lot of money]

Inclusion of verb forms in illustrative examples. In addition to their sensitivity to the wording of the context versus the wording of MWLD examples, participants were also sensitive to verb forms in examples. They firmly advocated for examples that would contain a variety of verb forms. They certainly preferred to see in the dictionary the specific verb form that appeared in their own context.

Use of boldface and italics in illustrative examples. Our students were quite sensitive to the use of boldface and italics. Students found boldface and italics frustrating, because they had no idea why one or the other was used. At times, students expressed appreciation for boldface, because for them it made some information more prominent (see also Herbst (1996) and Dziemianko (2014) on highlighting collocations in bold). On the other hand, among the students who commented on this issue, all were unanimous in their dislike for italics.

Discussion

Our participants point out that, while dictionaries are created by experts, they can be improved if dictionary makers “can learn something about average users;” if linguists can “think like the average person.” Below are the most important recommendations of the participants for the makers of learner’s dictionaries.

Improving look-up ability and choice of appropriate sense/definition

The participants paid little attention to part of speech and as a result lost time in locating in the dictionary the infrequent senses corresponding to the words in their contexts. During the course of the look-up activities, some participants noticed (and used) the MWLD’s small drop-down menu with part of speech labeled—and others never did. Those who did use this menu, often at a late point during their session, said that it made look-up easier.

The drop-down menu as it is currently presented in the online MWLD is not as salient, and thus not as useful as it could be in guiding users to the right part of speech. The menu needs to be of manageable length, to contain a manageable amount of material, and the right type of material. Among the target words of this study, the extreme case in the MWLD is the treatment of *mean*, a word with noun, verb, and adjective senses. The first three items visible in the seventeen-item drop-down menu are: *mean* (verb), *mean* (adjective), and *mean* (adjective). This gives no useful information that would allow a rational choice between the two adjective listings. Moreover, in this seventeen-item menu, only three items are visible at a time without scrolling. All of the words in the drop-down menu should be visible, not just the first three; to this end, the scroll bar on the right side of the drop-down menu should be eliminated.

As was noted above, the form of the target word in the context was the main driver in the participants’ efforts to locate the correct meaning in the dictionary—and that form often led them astray. In the main look-up box at the top of the MWLD screen, the users usually typed in the target word in exactly the same form that they found in their context.

Improving illustrative examples

Our subjects wanted full-sentence illustrative examples, even in those cases where lexicographers might feel a phrase or a brief word combination is adequate. While Atkins and Rundell (2008) maintain that balance in the amount of context in illustrative examples is something to strive for, our participants appeared to be saying that it would be better to err on the side of longer rather than shorter examples.

The subjects’ desire to have *the* example most closely related to their context is related to subjects’ interest in a greater number of examples. If the dictionary includes more examples, then learners have more of a possibility to find *the* example that works for them. If the dictionary offers more examples, then more forms

such as *fixing* and *taxed* can be represented within the full gamut of examples provided; this was another clear desire of our subjects.

While these learners said that they wanted more information from examples, they also said they do not want “TMI,” too much information. One subject said s/he would like to have more examples and longer examples, but not everywhere. Another subject, who was familiar with information technology design, suggested the possibility of clicking to access more examples or of clicking on a specific example to lengthen it into a full sentence or even short paragraph. This subject said that dictionary users should have the means “to go deeper” into the information, but that those who do not wish to do so should not be “force[d] ... to read everything.”

Improving presentation of dictionary information

One crystal-clear message from our participants concerned their dislike of italics: They have no idea what they mean or why dictionaries have them instead of boldface; they render text “unnoticeable;” they are “old fashioned.” The participants liked boldface, though they were not sure what it meant, either: One subject thought that boldface was the dictionary’s way of “telling the reader which [illustrative] example is more appropriate.”

One participant suggested the use of color to clarify the presentation of information: Different types of information could be in different colors. This idea was used by the developers of the print edition of the *Merriam–Webster’s Advanced Learner’s English Dictionary* (Perrault, 2008). In that dictionary, the illustrative examples appear to be both in boldface and in blue color. It was a great loss that the use of blue in the printed edition did not transfer to the MWLD online (the color of the online illustrative examples is black). Bringing blue back into this dictionary would help learners distinguish between the definition and the illustrative examples; adding other colors could further improve the look-up process.

Conclusion

The participants in this study enjoyed using the online *Merriam–Webster Learner’s Dictionary* and some of them planned to continue using it. At the same time, these subjects were forthcoming about the challenges they faced during the look-up activities they undertook. The perceptions of these advanced English learners are valuable to those seeking to improve the content as well as the format and presentation of information in learner’s dictionaries. While certainly it is difficult to display information in a way that is clear to a majority of learners, we cannot but agree with the statement of one of our participants: “You can be a great scientist but you need to convey information so that normal people can have access.”

Acknowledgements

The authors acknowledge the project, *Dictionary User Groups: What They Can Teach Lexicographers* (grant number: BI-US/17-18-033), which was financially supported by the Slovenian Research Agency. They also acknowledge the approval (2 March 2018) of the New Jersey City University (NJCU) Institutional Review Board for the Protection of Human Participants in Research, and the approval (5 March 2018) of the Ethics Committee of the Faculty of Arts, University of Ljubljana. Donna Farina thanks NJCU for travel support to Ljubljana, Slovenia. The authors thank NJCU, in particular Tamara Cunningham, Assistant Vice President for Global Initiatives, for providing housing and hospitality to Alenka Vrbinc and Marjeta Vrbinc during a research visit to the United States. Finally, the authors are very grateful to the study participants from undergraduate programs in the Faculty of Economics at the University of Ljubljana.

References

- Assigning CEFR ratings to ACTFL assessments. n.d. Retrieved from https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf
- Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.
- Bitenc Peharc, S., & Tratnik, A. (2014). *Umestitev nacionalnih izpitov iz angleščine v skupni evropski jezikovni okvir. Zaključno poročilo*. Retrieved from <https://www.ric.si/mma/SEJO%20Umestitev%20nacionalnih%20izpitov%20iz%20angle%20%20%20%20ine%20/2014073109411981/>
- Dziemińko, A. (2014). On the presentation and placement of collocations in monolingual English learners' dictionaries: Insights into encoding and retention. *International Journal of Lexicography*, 27(3), 259–279. doi.org/10.1093/ijl/ecu012.
- Hatherall, G. (1984). Studying dictionary use: Some findings and proposals. In R.R.K. Hartmann (Ed.). *LEXexter '83 Proceedings* (pp. 183–189). Tübingen: Max Niemeyer.
- Herbst, T. (1996). On the way to the perfect learners' dictionary: A first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. *International Journal of Lexicography*, 9(4), 321–357. doi.org/10.1093/ijl/9.4.321.
- Ilc, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*, 32(4), 443–462. doi.org/10.1177/0265532214562098.
- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage Publications.
- Lew, R., & de Schryver, G. M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), 341–359. doi:10.1093/ijl/ecu011.
- Lorentzen, H., & Theilgaard, L. (2012). Online dictionaries – How do users find them and what do they do once they have?. In R.V. Fjeld, & J.M. Torjusen (Eds). *Proceedings of the 15th EURALEX International Congress* (pp. 654–660). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
- Merriam–Webster learner's dictionary*. Retrieved from <http://learnersdictionary.com/> (MWLD)
- Perrault, S. J. (Ed). (2008). *Merriam–Webster's advanced learner's English dictionary*. Springfield, Massachusetts: Merriam–Webster, Incorporated.
- Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative Research in Accounting & Management*, 8(3), 238–264. doi.org/10.1108/11766091111162070.
- Učni načrt Angleščina. 2008. Retrieved from http://www.mss.gov.si/fileadmin/mss.gov.si/pageuploads/podrocje/ss/programi/2008/Gimnazije/UN_AN_GLESCINA_gimn.pdf
- Wingate, U. (2002). *The effectiveness of different learner dictionaries*. Tübingen: Max Niemeyer.

WESTERN LOANWORDS DERIVED WITH TURKISH SUFFIXES IN TURKISH DICTIONARIES

Fatih Doğru

Eskişehir Osmangazi University

Abstract

The Turkish language also interacts with some other languages and some of the reasons are geographical area, commercial relations, technological developments, the expansion of communication facilities, and social factors.

There are many foreign words in the Turkish language. Most of these words have a Western origin. These loanwords have become elements of the Turkish vocabulary throughout the centuries and some of them have been derived with some kind of Turkish suffixes.

In this study, Western loanwords derived with Turkish suffixes have been examined in Turkish dictionaries. According to the results of this study, these suffixes are derivational suffixes that include “suffixes that attach to nominals”. There are a small number of “suffixes that attach to verbs”.

The following languages were accepted as Western languages which are labeled as the origin of headwords in Türkçe Sözlük (Turkish Dictionary) (2015) and Türkçedeki Yabancı Sözcükler Sözlüğü (Dictionary of foreign words in Turkish) (2004): French, Italian, Greek, English, German, Spanish, Russian, Latin, Romaic, Armenian, Slavic, Bulgarian, Hungarian, Albanian, Portuguese, Norwegian, Germanic, Romanian, and the Australian native language.

The following Turkish suffixes that are attached to Western loanwords were designated in Turkish dictionaries, Türkçe Sözlük (2005) and Türkçedeki Yabancı Sözcükler Sözlüğü (2004): +Al, +CA, +CI⁴, +CI⁴k, +(a)cık, +CI⁴kI⁴, +CI⁴I, +CI⁴IAr, +CI⁴II⁴k, +DA, +DAn, +dAş, -gI⁴, +giller, -I⁴cI⁴, -I⁴ImAk, +(I⁴)m, +(I⁴)msI⁴, +(I⁴)mtrak, +I⁴ylA, +k, +la-Ø, +Iaç, +IAmA, +IAmAçI, +IAmAçIIIk, +IAmAk, +IAAnAbilir, +IAAndIrIlmA, +IAAndIrIlmAk, +IAAndIrmA, +IAAndIrmAk, +IAAnIş, +IAAnmA, +IAAnmAk, +IAAr, +IAArcA, +IAşmA, +IAşmAk, +IAştIrIcI, +IAştIrIcIIIk, +IAştIrIlmA, +IAştIrIlmAk, +IAştIrmA, +IAştIrmAk, +IAtIlmAk, +IAtIş, +IAtmA, +IAtmAk, +IAttIrmA, +IAttIrmAk, +IAyI⁴cI⁴, +IAyIş, +II⁴, +II⁴k, +II⁴IAr, +II⁴IAşmA, +II⁴IAşmAk, +II⁴II⁴k, -mA, -mAk, +sAl, +sAmA, +sAmAll, +sI⁴, +sI⁴IAr, +sI⁴z, +sI⁴zIAAndIrmAk, +sI⁴zIAr, +sI⁴zIAşmA, +sI⁴zIAşmAk, +sI⁴zIAştIrmA, +sI⁴zII⁴k, +ş.

Key Words: Turkish dictionaries, language relations, Western loanwords, Turkish suffixes

1. Introduction

The processes of word formation include derivation, compounding, acronymy, borrowing, antonomasia, conversion, blending, backformation, clipping, lexical abbreviation, folk etymology, and coinage (Hartmann and James, 1998: 156; Burkhanov, 1998: 264; Nasser, 2008: 73). The most frequently quoted

word formation processes are derivation and compounding (Burkhanov, 1998: 264). Plag (2002: 22) classified the derivation in two parts: affixation and non-affixation. Affixation has also three parts: prefixation, suffixation, and infixation. This study is focused on suffixation. In Turkish, considering its characteristics, the most common word formation method is derivation by suffixation. Göksel and Kerslake (2005: 51) states that derivation is the creation of a new lexical item and the vast majority of derivation in Turkish is achieved through suffixation. When a derivational suffix attaches to a stem it produces a new word connected in meaning to that stem. Some derivational suffixes change the class of the word they attach to (Göksel and Kerslake, 2005: 51). As in the Turkish origin words, words of foreign origin can also be lexicalized by the way of derivation in Turkish. A Turkish derivational suffix can be added to a foreign origin root. This root can be Eastern origin like Arabic, and Persian, etc. or it can be the Western origin. In addition to words of Eastern origin, a large number of Western origin words have been lexicalized by this method in Turkish. The term "Western" is a cultural expression rather than a geographical direction. There are political and cultural connotations of the West in Turkish history. In connection with the European enlightenment process, the Western concept in Turkish is also included in the modernization. Otherwise in "Türkçede Batı Kökenli Kelimeler Sözlüğü" (The Dictionary of Western origin words), the Greek and Armenian loanwords, which have been used in the Turkish language since the ancient times, are also included (Akalin et. al., 2015: 6). In this study, the words accepted as Western loanwords as stated in this dictionary were accepted as Western loanwords. The following languages were accepted as Western languages which are labeled the origin of headwords in "Türkçe Sözlük" (Turkish Dictionary) (2015) and "Türkçedeki Yabancı Sözcükler Sözlüğü" (Dictionary of foreign words in Turkish) (2004): French, Italian, Greek, English, German, Spanish, Russian, Latin, Romaic, Armenian, Slavic, Bulgarian, Hungarian, Albanian, Portuguese, Norwegian, Germanic, Romanian, and the Australian native language. The origin of some words is unspecified and the headwords of some roots do not exist in the dictionaries.

The research questions of this study are as follows. Which Turkish suffixes were added to Western loanwords? Which Western loanwords derived with Turkish suffixes were lexicalized? What are the basic functions of the Turkish suffixes which were added to Western loanwords in the Turkish dictionaries?

The aim of this study is to identify the Turkish suffixes that were added to Western loanwords in Turkish dictionaries and to reveal their basic functions. Thus, derivation which is one of the word formation methods of Turkish will be designated in the Western loanwords. This study does not aim to explain the use reasons and the intensity of all Western loanwords. This information can be accessed from Sezgin (2004) and Sarı (2008).

2. Method

Each headword in printed dictionaries of Türkçe Sözlük (TS) and Türkçedeki Yabancı Sözcükler Sözlüğü (TYSS) was checked individually. The words identified as loanwords from Western languages are detected, the headword forms with a Turkish suffix of these loanwords in the examined dictionaries are determined, and the basic functions of the suffixes are designated.

In the results section, the outcomes of TS are given constitutively. If a Western loanword derived with a Turkish suffix that is not in the TS but is in the TYSS, then its outcomes are included. This method was applied to prevent a repetition of total numbers.

3. Results

3.1. Languages labeled as a Western language in TS and TYSS and words derived from these languages

There are 63,818 headwords in the TS. 3426 of them (5,368%) were created by attaching them with Turkish suffix(es) to Western loanwords. These headwords were derived from 1523 roots. There are 487 words derived from 352 Western roots in the TYSS which are not in the TS.

	French loanwords	French loanwords derived with Turkish suffix
TS	801	1887
TYSS (not in TS)	177	239
Total	978	2126

Table 1. French loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 1, in TS, there are 1887 headwords derived with Turkish suffixes from 801 French labeled headwords. In TYSS, there are 239 headwords derived with Turkish suffixes from 177 French labeled headwords that are not included in the TS.

	Italian loanwords	Italian loanwords derived with Turkish suffix
TS	215	545
TYSS (not in TS)	48	72
Total	263	617

Table 2. Italian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 2, in TS, there are 545 headwords derived with Turkish suffixes from 215 Italian labeled headwords. In TYSS, there are 72 headwords derived with Turkish suffixes from 48 Italian labeled headwords that are not included in the TS.

	Greek loanwords	Greek loanwords derived with Turkish suffix
TS	123	385
TYSS (not in TS)	40	73
Total	163	458

Table 3. Greek loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 3, in TS, there are 385 headwords derived with Turkish suffixes from 123 Greek labeled headwords. In TYSS, there are 73 headwords derived with Turkish suffixes from 40 Greek labeled headwords that are not included in the TS.

	English loanwords	English loanwords derived with Turkish suffix
TS	91	162
TYSS (not in TS)	26	32
Total	117	194

Table 4. English loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 4, in TS, there are 162 headwords derived with Turkish suffixes from 91 English labeled headwords. In TYSS, there are 32 headwords derived with Turkish suffixes from 26 English labeled headwords that are not included in the TS.

	German loanwords	German loanwords derived with Turkish suffix
TS	14	32
TYSS (not in TS)	3	6
Total	17	38

Table 5. German loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 5, in TS, there are 32 headwords derived with Turkish suffixes from 14 German labeled headwords. In TYSS, there are 6 headwords derived with Turkish suffixes from 3 German labeled headwords that are not included in the TS.

	Spanish loanwords	Spanish loanwords derived with Turkish suffix
TS	10	21
TYSS (not in TS)	4	4
Total	14	25

Table 6. Spanish loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 6, in TS, there are 21 headwords derived with Turkish suffixes from 10 Spanish labeled headwords. In TYSS, there are 4 headwords derived with Turkish suffixes from 4 Spanish labeled headwords that are not included in the TS.

	Armenian loanwords	Armenian loanwords derived with Turkish suffix
TS	8	16
TYSS (not in TS)	3	4
Total	11	20

Table 7. Armenian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 7, in TS, there are 16 headwords derived with Turkish suffixes from 8 Armenian labeled headwords. In TYSS, there are 4 headwords derived with Turkish suffixes from 3 Armenian labeled headwords that are not included in the TS.

	Russian loanwords	Russian loanwords derived with Turkish suffix
TS	9	14
TYSS (not in TS)	-	-
Total	9	14

Table 8. Russian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 8, in TS, there are 14 headwords derived with Turkish suffixes from 9 Russian labeled headwords.

	Romaic loanwords	Romaic loanwords derived with Turkish suffix
TS	9	12
TYSS (not in TS)	-	-
Total	9	12

Table 9. Romaic loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 9, in In TS, there are 12 headwords derived with Turkish suffixes from 4 Romaic labeled headwords.

	Slavic loanwords	Slavic loanwords derived with Turkish suffix
TS	6	10
TYSS (not in TS)	1	1
Total	7	11

Table 10. Slavic loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 10, in TS, there are 10 headwords derived with Turkish suffixes from 6 Slavic labeled headwords. In TYSS, there are 1 headwords derived with Turkish suffixes from 1 Slavic labeled headwords that are not included in the TS.

	Latin loanwords	Latin loanwords derived with Turkish suffix
TS	7	9
TYSS (not in TS)	3	4
Total	10	13

Table 11. Latin loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 11, in In TS, there are 9 headwords derived with Turkish suffixes from 7 Latin labeled headwords. In TYSS, there are 4 headwords derived with Turkish suffixes from 3 Latin labeled headwords that are not included in the TS.

	Hungarian loanwords	Hungarian loanwords derived with Turkish suffix
TS	2	9
TYSS (not in TS)	1	3
Total	3	12

Table 12. Hungarian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 12, in TS, there are 9 headwords derived with Turkish suffixes from 2 Hungarian labeled headwords. In TYSS, there are 3 headwords derived with Turkish suffixes from 1 Hungarian labeled headwords that are not included in the TS.

	Bulgarian loanwords	Bulgarian loanwords derived with Turkish suffix
TS	3	7
TYSS (not in TS)	-	-
Total	3	7

Table 13. Bulgarian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 13, in TS, there are 7 headwords derived with Turkish suffixes from 3 Bulgarian labeled headwords.

	Portuguese loanwords	Portuguese loanwords derived with Turkish suffix
TS	2	3
TYSS (not in TS)	-	-

Total	2	3
--------------	---	---

Table 14. Portuguese loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 14, in TS, there are 3 headwords derived with Turkish suffixes from 2 Portuguese labeled headwords.

	Albanian loanwords	Albanian loanwords derived with Turkish suffix
TS	1	1
TYSS (not in TS)	-	-
Total	1	1

Table 15. Albanian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 15, in TS, there are 1 headwords derived with Turkish suffixes from 1 Albanian labeled headwords.

	Norwegian loanwords	Norwegian loanwords derived with Turkish suffix
TS	1	1
TYSS (not in TS)	-	-
Total	1	1

Table 16. Norwegian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 16, in TS, there are 1 headwords derived with Turkish suffixes from 1 Norwegian labeled headwords.

	Turkish+French roots	Turkish+French derived with Turkish suffix
TS	8	9
TYSS (not in TS)	-	-

Total	8	9
--------------	---	---

Table 17. Turkish+French roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 17, in TS, there are 9 headwords derived with Turkish suffixes from 8 Turkish+French labeled compound words.

	Turkish+Greek roots	Turkish+Greek derived with Turkish suffix
TS	3	3
TYSS (not in TS)	2	2
Total	5	5

Table 18. Turkish+Greek roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 18, in TS, there are 3 headwords derived with Turkish suffixes from 3 Turkish+Greek labeled compound words. In TYSS, there are 2 headwords derived with Turkish suffixes from 2 Turkish+Greek labeled compound words that are not included in the TS.

	Turkish+Italian roots	Turkish+Italian derived with Turkish suffix
TS	2	3
TYSS (not in TS)	1	1
Total	3	4

Table 19. Turkish+Italian roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 19, in TS, there are 3 headwords derived with Turkish suffixes from 2 Turkish+Italian labeled compound words. In TYSS, there are 1 headwords derived with Turkish suffixes from 1 Turkish+Italian labeled compound words that are not included in the TS.

	French+English loanwords	French+English derived with Turkish suffix
TS	2	3

TYSS (not in TS)	-	-
Total	2	3

Table 20. French+English loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 20, in TS, there are 3 headwords derived with Turkish suffixes from 2 French+English labeled compound words.

	Italian+Persian loanwords	Italian+Persian derived with Turkish suffix
TS	2	3
TYSS (not in TS)	-	-
Total	2	3

Table 21. Italian+Persian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 21, in TS, there are 3 headwords derived with Turkish suffixes from 2 Italian+Persian labeled compound words.

	French+Arabic loanwords	French+Arabic derived with Turkish suffix
TS	2	2
TYSS (not in TS)	1	1
Total	3	3

Table 22. French+Arabic loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 22, in TS, there are 2 headwords derived with Turkish suffixes from 2 French+Arabic labeled compound words. In TYSS, there are 1 headwords derived with Turkish suffixes from 1 French+Arabic labeled compound words that are not included in the TS.

	French+Turkish roots	French+Turkish derived with Turkish suffix
--	-----------------------------	---

TS	2	2
TYSS (not in TS)	5	5
Total	7	7

Table 23. French+Turkish roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 23, in TS, there are 2 headwords derived with Turkish suffixes from 2 French+Turkish labeled compound words. In TYSS, there are 5 headwords derived with Turkish suffixes from 5 French+Turkish labeled compound words that are not included in the TS.

	Turkish+Latin roots	Turkish+Latin derived with Turkish suffix
TS	1	1
TYSS (not in TS)	-	-
Total	1	1

Table 24. Turkish+Latin roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 24, in TS, there are 1 headwords derived with Turkish suffixes from 1 Turkish+Latin labeled compound words.

	Turkish+English roots	Turkish+English derived with Turkish suffix
TS	1	1
TYSS (not in TS)	-	-
Total	1	1

Table 25. Turkish+English roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 25, in TS, there are 1 headwords derived with Turkish suffixes from 1 Turkish+English labeled compound words.

	French+Persian loanwords	French+Persian derived with Turkish suffix
TS	1	1
TYSS (not in TS)	-	-
Total	1	1

Table 26. French+Persian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 26, in TS, there are 1 headwords derived with Turkish suffixes from 1 French+Persian labeled compound words.

	Germanic loanwords	Germanic loanwords derived with Turkish suffix
TS	-	-
TYSS (not in TS)	2	3
Total	2	3

Table 27. Germanic loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 27, in TYSS, there are 3 headwords derived with Turkish suffixes from 2 Germanic labeled headwords that are not included in the TS.

	Romanian loanwords	Romanian loanwords derived with Turkish suffix
TS	-	-
TYSS (not in TS)	1	1
Total	1	1

Table 28. Romanian loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 28, in TYSS, there are 1 headwords derived with Turkish suffixes from 1 Romanian labeled headwords that are not included in the TS.

	Greek+Turkish roots	Greek+Turkish derived with Turkish suffix
TS	-	-
TYSS (not in TS)	3	3
Total	3	3

Table 29. Greek+Turkish roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 29, in TYSS, there are 3 headwords derived with Turkish suffixes from 3 Greek+Turkish labeled compound words that are not included in the TS.

	Greek+Arabic roots	Greek+Arabic derived with Turkish suffix
TS	-	-
TYSS (not in TS)	1	2
Total	1	2

Table 30. Greek+Arabic roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 30, in TYSS, there are 2 headwords derived with Turkish suffixes from 1 Greek+Arabic labeled compound words that are not included in the TS.

	Australian native language loanwords	Australian native language loanwords derived with Turkish suffix
TS	-	-
TYSS (not in TS)	2	2
Total	2	2

Table 31. Australian native language loanwords and words derived with Turkish suffixes in TS and TYSS

As shown in Table 31, in TYSS, there are 2 headwords derived with Turkish suffixes from 2 Australian native language labeled headwords that are not included in the TS.

	Unspecified origin (UnO.)	Unspecified origin derived with Turkish suffix
TS	46	92
TYSS (not in TS)	16	25
Total	62	117

Table 32. Unspecified origin (UnO.) roots and words derived with Turkish suffixes in TS and TYSS

As shown in Table 32, in TS, there are 92 headwords derived with Turkish suffixes from 46 headwords which have unspecified origins. In TYSS, there are 25 headwords derived with Turkish suffixes from 16 headwords which have unspecified origins that are not included in the TS.

	Roots which do not exist in the dictionary (RNED)	Words derived with Turkish suffix whose root does not exist in the dictionary
TS	141	179
TYSS (not in TS)	8	14
Total	149	193

Table 33. Roots which do not exist in the dictionary (RNED) and words derived with Turkish suffix whose root does not exist in the dictionary in TS and TYSS

As shown in Table 33, in TS, there are 179 headwords derived with Turkish suffixes from 141 roots that do not exist in the dictionary. In TYSS, there are 14 headwords derived with Turkish suffixes from 8 roots that do not exist in the dictionary.

3.2. Turkish suffixes that attach to Western loanwords

3.2.1. Suffixes that attach to nominals

Suffixes that attach to nominals create both verbs and other nominals (nouns, adjectives, and adverbs) (Göksel and Kerslake, 2005: 56).

3.2.1.1. Suffixes that attach to nominals to form verbs

The following Turkish suffixes that attach to nominals to form verbs are attached to Western loanwords in TS and TYSS:

+IAmAk

afişlemek (French), aforozlamak (Greek), vakumlamak (Latin) ... (TS, 2005).

tokalamak (Italian), matizlemek (Greek), paspaslamak (French) ... (TYSS, 2004).

+IAnmAk

aforozlanmak (Yun), streslenmek (English), pompalanmak (Italian) ... (TS, 2005).

nötrlenmek (French), kokorozlanmak (Italian), formatlanmak (English) ... (TYSS, 2004).

+IAndIrmAk

gruplandırmak (French), kadrolandırmak (Italian), kuruluşlandırmak (German) ... (TS, 2005).

kadrolandırmak (Italian) (TYSS, 2004).

+IAndIrIlmAk

projelendirilmek (French), kristallendirilmek (French) ... (TS, 2005).

kredilendirilmek (French), elektrikleştirilmek (French), sınırlandırılmak (Greek) ... (TYSS, 2004).

+IAşmAk

sendikalaşmak (French), tokalaşmak (Italian), moruklaşmak (Armenian) ... (TS, 2005).

kanserleşmek (French), kadrolaşmak (Italian), Frenkleşmek (Germanic) ... (TYSS, 2004).

+IAştIrmAk

şıklıştırmak (French), sloganlaştırmak (English), çeteleştirmek (Bulgarian) ... (TS, 2005).

kangrenleştirmek (French), abanozlaştırmak (Greek), koklaştırmak (English) ... (TYSS, 2004).

+IAştIrIlmAk

laikleştirilmek (French), sembolleştirilmek (French), polimerleştirilmek (French) (TS, 2005).

mitleştirilmek (French), Hıristiyanlaştırılmak (Greek), koklaştırılmak (English) ... (TYSS, 2004).

+IAtmAk

kodlatmak (French), çapalamak (Italian), demetletmek (Greek) ... (TS, 2005).

presletmek (French), kasalamak (Italian), fertikletmek (German) ... (TYSS, 2004).

+IAtIlmAk

paketletilmek (French), bombalatılmak (Italian), kundaklatılmak (Greek) ... (TYSS, 2004).

+IAttIrmAk

kaşeleştirmek (French) (TS, 2005).

+II⁴IAşmAk

Amerikalılařmak (UnO.), Avrupalılařmak (RNED) (TS, 2005).

+sI⁴zlAřmAk

randımansızlařmak (French), vizyonsuzlařmak (French) (TYSS, 2004).

+sI⁴zlAndIrmAk

mikropsuzlandırmak (French), (TYSS, 2004).

+la-Ø

kopyalayayıřtır (Italian) (TS, 2005).

+lAnAbilir

oksitlenebilir (French) (TYSS, 2004).

3.2.1.2. Suffixes that attach to nominals to form nominals

The following Turkish suffixes that attach to nominals to form nominals are attached to Western loanwords in TS and TYSS:

+Al

erosal (French), helisel (French) (TS, 2005).

+CA

amatörce (French), kalantorca (Italian), anavakça (Armenian) ... (TS, 2005).

hoyratça (Greek), snopça (English), lümpence (German) ... (TYSS, 2004).

+CI⁴

abajurcu (French), sambacı (Portuguese), slalomcu (Norwegian) ... (TS, 2005).

izmaritçi (Greek), tolkşovcu (English), kanserbilimci (French+Turkish) ... (TYSS, 2004).

+CI⁴k

kanalcık (French), pusulacık (Italian), filizcik (Greek) ... (TS, 2005).

+CI⁴II⁴k

afiřçilik (French), řapkacılık (Russian), sobacılık (Hungarian) ... (TS, 2005).

kambiyoculuk (Italian), fındıkçılık (Greek), basketçilik (English) ... (TYSS, 2004).

+(a)cık

minnacık (French) (TS, 2005).

+CI⁴kII⁴

kanalcıklı (French) (TS, 2005).

+CII

silisçil (French), larvacıl (Latin) (TS, 2005).

+CI⁴Ar

kalafatçılar (Italian) (TS, 2005).

+DA

pratikte (French) (TS, 2005).

+DAn

bodoslamadan (Greek), anafordan (Greek), avantadan (UnO.) (TS, 2005).

koftiden (Greek), plaçkadan (Albanian) (TYSS, 2004).

+dAş

sınırdaş (Greek) (TS, 2005).

+giller

kangurugiller (French), fasulyegiller (Greek), iguanagiller (Spanish) ... (TS, 2005).

kaktüsgiller (French), papağangiller (Italian), lapinagiller (Greek) ... (TYSS, 2004).

+(I⁴)m

efendim (Greek) (TS, 2005).

+(I⁴)msI⁴

grimsi (French), limonumsu (Greek), patatesimsi (Romaic) ... (TS, 2005).

diskimsi (French), sarkomsu (French), kartonumsu (French) (TYSS, 2004).

+(I⁴)mtırak

grimtrak (French), bordomtrak (French) (TS, 2005).

+I⁴yla

kanalıyla (French) (TS, 2005).

tomarla (Greek) (TYSS, 2004).

+k

minik (French) (TS, 2005).

+lAç

indükleç (RNED) (TS, 2005).

+lAmA

afişleme (French), aforozlama (Greek), röntgenleme (German) ... (TS, 2005).

hidratlama (French), milleme (Greek), indükleme (RNED -French) (TYSS, 2004).

+lAmAcl

planlamacı (French), sondajlamacı (French) (TS, 2005).

+lAmAcIIIk

planlamacılık (French), sondajlamacılık (French) (TS, 2005).

+lAndIrmA

puanlandırma (French), kadrolandırma (Italian), gümrüklendirme (Greek) ... (TS, 2005).

+lAndIrIlmA

projelendirilme (French), kristallendirilme (French) (TS, 2005).

+lAnIş

paketleniş (French), çapalanış (Italian), sınırlanış (Greek) ... (TS, 2005).

+lAnmA

postalanma (Italian), çerezlenme (Greek), tabakalanma (Spanish) ... (TS, 2005).

kalamınlenme (French) (TYSS, 2004).

+lAr

penguenler (French), balinalar (Italian), fundalar (Greek) ... (TS, 2005).

antrasitler (French), papağanlar (Italian), amfibyumlar (Greek) ... (TYSS, 2004).

+lArcA

milyarlarca (French), milyonlarca (French) (TS, 2005).

tonlarca (French) (TYSS, 2004).

+lAşmA

kristalleşme (French), makineleşme (Italian), paslaşma (English) ... (TS, 2005).

oksitleşme (French), mermerleşme (Greek) maltlaştırma (English) ... (TYSS, 2004).

+lAştIrIcI

mekanikleştirici (French), mineralleştirici (French), ozonlaştırıcı (French) (TS, 2005).

nitratlaştırıcı (French), eterleştirici (French) (TYSS, 2004).

+lAştIrIcIIIk

mekanikleştiricilik (French) (TS, 2005).

+lAştIrmA

globalleştirme (French), tiyatrolaştırma (Italian), sloganlaştırma (English) ... (TS, 2005).

maltlaştırma (English) (TYSS, 2004).

+lAştIrIlmA

laikleştirilme (French), sembolleştirilme (French), polimerleştirilme (French) (TS, 2005).

+lAtIş

demetletiş (Greek) (TS, 2005).

+lAtmA

badanalatma (French), kaskolatma (Italian), gübreletme (Greek) ... (TS, 2005).

+lAttIrmA

kaşelettirme (French) (TS, 2005).

+lAyI⁴cI⁴

ozonlayıcı (French), bantlayıcı (French), frenleyici (French) (TS, 2005).

sınırlayıcı (Greek) (TYSS, 2004).

+lAyIş

fırınlayış (Greek), kolalayış (Italian), paketleyiş (French) ... (TS, 2005).

+II⁴

abajurlu (French), aforozlu (Greek), sigaralı (Spanish) ... (TS, 2005).

vanilyalı (Spanish), flüorlu (Latin), vişneli (Slavic) ... (TYSS, 2004).

+II⁴k

abonelik (French), acentelik (Italian), krallık (Slavic) ... (TS, 2005).

pijamalık (French), kalantorluk (Italian), çaçalık (Greek) ... (TYSS, 2004).

+II⁴IAr

trakeliler (French), karinalılar (Italian), palamutlular (Greek) ... (TS, 2005).

tallılar (French) (TYSS, 2004).

+II⁴IAşmA

tonlulaşma (French), Amerikalılaşma (UnO.), Avrupailulaşma (RNED) (TS, 2005).

+II⁴II⁴k

kültürlülük (French), sigortalılık (Italian), Avrupailuluk (RNED) ... (TS, 2005).

+sAI

fiziksel (French), silindirsel (French), biyokimyasal (French+Arabic) ... (TS, 2005).

fotokimyasal (French), atombilimsel (French+Turkish), fotokimyasal (French+Arabic) ... (TYSS, 2004).

+sAmA

alaysama (Greek) (TYSS, 2004).

+sAmAll

alaysamalı (Greek) (TYSS, 2004).

+sI⁴

betonsu (French), Amerikansı (English), alaysı (Romaic) ... (TS, 2005).

+sI⁴IAr

pelikansılar (French) (TS, 2005).

+sI⁴z

sansürsüz (French), faturasız (Italian), şapkasız (Russian) ... (TS, 2005).

risksiz (French), kadrosuz (Italian), kilitsiz (Greek) ... (TYSS, 2004).

+sI⁴zIAr

bloksuzlar (French) (TYSS, 2004).

+sI⁴zIAşmA

tonsuzlaşma (French), akortsuzlaşma (French) (TS, 2005).

+sI⁴zIAştIrmA

akortsuzlaştırma (French) (TS, 2005).

+sI⁴zII⁴k

şanssızlık (French), golsüzlük (English), temelsizlik (Greek) ... (TS, 2005).

randımansızlık (French) (TYSS, 2004).

+ş

minnoş (French) (TYSS, 2004).

3.2.2. Suffixes that attach to verbs

Suffixes that attach to verbs create new words which are either nominals (noun, adjective or adverb) or verbs (Göksel and Kerslake, 2005: 52). There are a small number of examples in TS and TYSS.

3.2.2.1. Suffixes that attach to verbs to form verbs

The following Turkish suffixes that attach to verbs to form verbs are attached to Western loanwords in TS and TYSS:

-I⁴ImAk

polarılmak (French) (TYSS, 2004).

3.2.2.2. Suffixes that attach to verbs to form nominals

The following Turkish suffixes that attach to verbs to form nominals are attached to Western loanwords in TS and TYSS:

-gI⁴

polargı (French) (TS, 2005).

-I⁴cI⁴

polarıcı (French) (TS, 2005).

-mA

polarma (French) (TS, 2005).

-mAk

polarmak (French) (TS, 2005).

manyamak (French) (TYSS, 2004).

4. Discussion and Conclusion

There are 63,818 headwords in TS and 3426 of them (5,368%) were created by attaching them with Turkish suffix or suffixes to Western loanwords.

In TS and TYSS, the most common Western language is French whose words were derived with Turkish suffixes. In TS, there are 1887 headwords derived with Turkish suffixes from 801 French root headwords. In TYSS, there are 239 headwords derived with Turkish suffixes from 177 French root headwords that are not included in the TS.

Turkish suffixes are most commonly attached to the following Western loanwords: “sınır” (Greek), “akort” (French) and “plan” (French).

“**sınır**” (*Gre.*) (13 single-words + 8 multi-words): sınırdaş, sınırdaşlık, sınırlama, sınırlamak, sınırlandırma, sınırlandırmak, sınırlanış, sınırlanma, sınırlanmak, sınırlayış, sınırlı (sınırlı doğru, sınırlı ortaklık, sınırlı sayı, sınırlı sorumluluk), sınırsız (sınırsız doğru, sınırsız sayı, sınırsız sorumluluk, sınırsız yetki), sınırsızlık

“**akort**” (*Fr.*) (13): akortçu, akortlama, akortlamak, akortlanma, akortlanmak, akortlatma, akortlatmak, akortlu, akortsuz, akortsuzlaşma, akortsuzlaşmak, akortsuzlaştırma, akortuzlaştırmak

“**plan**” (*Fr.*) (12 single-words + 3 multi-words): plancı, plancılık, planlama, planlamacı, planlamacılık, planlamak, planlanış, planlanma, planlanmak, planlı (planlı büyüme, planlı ekonomi), plansız (plansız programsız), plansızlık.

The most common Turkish suffixes are +**CI⁴** (559), +**II⁴** (535), +**II⁴k**, (464), and +**CI⁴II⁴k** (409) which are attached to Western loanwords. All these common suffixes derived nouns from nouns.

Future Work

Turkish suffixes which are attached to Western loanwords will be classified according to their roots. Their functions will be discussed and the list of the words derived with these suffixes will be prepared.

References

Burkhanov, I. (1998). *Lexicography: A dictionary of basic terminology*. Wydawn. Wyższej Szkoły Pedagogicznej w Rzeszowie.

- Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar*. London and NewYork: Routledge.
- Hartmann R.R.K. & James, G. (1998). *Dictionary of lexicography*. London and NewYork: Routledge.
- Nasser, M. (2008). "Processes of word formation in English and Arabic", *Journal of the College of Education*, 2/3, 71-87.
- Plag, I. (2002). *Word-formation in English*. Cambridge University Press.
- Pusküllüoğlu, A. (2004). *Türkçedeki Yabancı Sözcükler Sözlüğü* (Fifth edition). Ankara: Arkadaş Press.
- Sarı, M. (2008). *Türkçenin Batı Dilleriyle İlişkisi*. Ankara: Türk Dil Kurumu Press.
- Sezgin, F. (2004). *Türkçede Batı Kaynaklı Kelimelerin Yoğunluğu*. Ankara: Türk Dil Kurumu Press.
- Türkçe Sözlük* (2005). Ankara: Türk Dil Kurumu Press.
- Türkçede Batı Kökenli Kelimeler Sözlüğü* (2015). Edited by Şükrü Halûk Akalın... [et Al.]. Ankara: Türk Dil Kurumu Press.

A BIBLIOMETRIC STUDY OF LEXIKOS

Ferdi Bozkurt

Anadolu University

Abstract

There are hundreds of academic journals in the field of science. These journals have published many scientific studies on various fields of linguistics. On the other hand there are several international academic journals published in the field of lexicography. The International Journal of Lexicography, Dictionaries, Lexicography Journal of ASIALEX, and Lexikos are the most prominent ones. More and more academic articles, reviews, contemplative articles, research articles, project(s) related to lexicography are published in these journals. These lexicographical texts are expanding the literature of lexicography.

Bibliometric methods that were firstly used by Pritchard are able to provide various details to the science field in order to demonstrate the significant points and trends in academic journals more clearly. To reveal which publications are most cited, which regions and countries are the most contributive ones, which authors are most producing, and which authors have more influential scientifically is a meaningful study for lexicography. Gilles-Maurice de Schryver has conducted two bibliometric studies considering lexicography in the same year. Also, one of the studies was about the Lexikos journal which is the the

official voice of AFRILEX. De Schryver focused on the types of academic texts published by Lexikos from 1991 to 2018. He revealed important information about the texts such as the types of contributions, the authors' countries, the most cited authors, the amount of co-authored studies, and the most cited publications.

In the current study, 14 years of Lexikos journal will be analysed through using bibliometric methods. The aim of this study is to examine which publications are most cited, which regions and countries are most contributive, which authors are most producing, and which authors have most scientific influential.

Key Words: articles, bibliometrics, lexicography, Lexikos journal, lexicographic texts,

1. Introduction

One of the stages of scientific research is the publication process, during which the obtained data is transformed into information. According to Day (1995: ix) “The goal of scientific research is publication”. Most academies around the world regard scientific publications as the primary outputs of scientific research. Most institutions measure the success of academic researchers according to these outcomes. The Research Excellence Framework (REF), an organization that assesses the quality of research of higher education institutions in the UK, highlights the publications first while listing the research outputs of the report named Panel criteria and working methods published in 2012. On the report, types of publications are listed as follows: (REF, 2012):

- papers in peer-reviewed journals
- papers in conference proceedings
- research reports to government departments, charities, the voluntary sector, professional bodies, industry or commerce
- monographs
- books and book chapters
- editorial
- review

As can be seen in this order, “papers in peer-reviewed journals” were given in first place. The most important meeting point of papers in peer-reviewed journals, papers in conference proceedings, Editorial and reviews is undoubtedly the academic journals.

Today, there are tens of thousands of academic journals in the world. In the digital age we live in, there is no exact number of how many journals are published. The number of journals published by the largest publishers in the world are as follows:

- Elsevier 2,689 journals³⁸,
- Springer Link 3,568 journals³⁹,

³⁸ <https://www.elsevier.com/catalog?producttype=journals&cat0=&q=&imprintname=&categoryrestriction=&author=&sort=datedesc> Last Accessed March 04, 2019.

³⁹ <https://link.springer.com/search?facet-content-type=%22Journal%22> Last Accessed March 04, 2019.

- Taylor & Francis 2,700 journals⁴⁰,
- Wiley 1,600 journals⁴¹,
- SAGE Journals more than 1.000 journals⁴²,
- Oxford Academic publishes more than 200 journals⁴³.

In parallel with the fact that there are so many academic journals, the number of articles has reached very high numbers. For example, in the database of Springer Link, there are 6,723,383 articles and 1,084,592 conference papers⁴⁴.

2. Lexicographic Journals in the World

There are thousands of academic journals in the world that publish articles related to the various branches of science, and there are also journals that publish articles related to lexicography. The oldest of these journals is “*Dictionaries*”. This journal which belongs to The Dictionary Society of North America (DSNA) has published 39 volumes so far and has been running since 1979.

Second major journal, “The International Journal of Lexicography”, was launched in 1988 by EURALEX. Until May 2019, *The International Journal of Lexicography* has published 32 volumes and 125 issues. Another journal in the Springer publication group is *Lexicography Journal of ASIALEX*, owned by ASIALEX. This journal has been running since 2014 and has published 5 Volumes and 10 issues so far. The other important academic journal is “*Lexikos*”, which constitutes the center of this study. This journal was first published by the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) in 1991 and with the establishment of the African Association for Lexicography, (AFRILEX) has become the voice of this association in 1995. The journal has published 28 volumes by 2019.

3. Bibliometrics and Lexicography

Alan Pritchard (1969), who used the term bibliometrics for the first time, defined it as follow: “application of mathematics and statistical methods to books and other media of communication”. As can be seen from the definition, it is possible to use bibliometrics for statistical method in books and other media of communication. Based on this definition, this method has been used to make trend analysis of the subjects of articles published in different journals related to a branch of science (Thompson, 2018) and sometimes it has been used to reveal the trends of a journal over the years (Abdi et. al., 2018).

Various bibliometric studies have been carried out not only in various fields of linguistics but also in the field of lexicography (Arik, 2015; Mohsen et. al. 2017; Lei & Liu, 2018). Lei & Liao 2017). Schryver (2009a; 2009b) conducted two very important analysis in the field of lexicography. In his study, Schryver (2009a) analyzed one of the oldest journals in the field of lexicography, the *International Journal of Lexicography*, by taking it to the center. He also made comparisons with different disciplines (Linguistics and Applied Linguistics) by comparing lexicographic journals, (*Lexikos* and *Dictionary*) which are other important journals of lexicography and linguistics field. In the second study, Schryver (2009b) conducted

⁴⁰ <https://www.tandfonline.com/> Last Accessed March 04, 2019.

⁴¹ <https://onlinelibrary.wiley.com/library-info/products/journals> Last Accessed March 05, 2019.

⁴² <https://uk.sagepub.com/en-gb/eur/sage-journals> Last Accessed March 04, 2019.

⁴³ <https://academic.oup.com/journals> Last Accessed March 02, 2019.

⁴⁴ <https://link.springer.com/> Last Accessed May 02, 2019.

a research on the Lexikos journal and took an 18-year statistical photograph of this journal. However, no examination was made in the study on citations to the publications in Lexikos. Both of these studies are very important in terms of presenting contributions to the field of lexicography. Bibliometric studies in the field of lexicography are undoubtedly important to see the distance that this field of study has taken in time and to guide future studies. Making the statistical analysis of academic journals, which is probably the most important meeting point of academic studies such as articles, presentations, reviews etc. will guide the future scientific studies.

4. The Objective of the Study

The aim of this study is to reveal a research trend by examining the scientific texts published in Lexikos Journal between 2005 and 2018, by examining their types, the changes on the texts in the years, the types of publications of scientific texts, the number of citations to publications, the countries where the publications were conducted and the number of collaborations in publications. For this purpose, the records of 14 issues of Lexikos published between 2005 and 2018 were examined.

5. Research Questions

The current study will focus on the following research questions:

- What is the number of the text types in Lexikos and how is their distribution by year and the publishing language?
- How is the numerical distribution of the authors in terms of contribution?
- Who are the most prolific writers in the journal?
- How is the distribution of the countries or regions, affiliations of the contributing authors?
- What is the impact of Lexikos Journal in terms of citation?
- What are the most cited texts in the journal
- Which authors are most cited?

6. Methodology

Although Lexikos published its first volume in 1991, it entered the Social Science Citation Index in 2005. Therefore, the articles published by the journal before 2005 are not included in the Web of Science Core Collection. Web of Science Core Collection (<http://apps.webofknowledge.com>) was used as the database of this study. In this study, data related to each other is presented with descriptive and explanatory methods.

7. Findings and Interpretation

LXIKOS Journal has been included in the Thomson Reuters Web of Science Citation Index since 2005. LEXIKOS Journal is indexed in Social Sciences Citation Index, Arts & Humanities Citation Index, Current Contents - Social & Behavioral Sciences and Current Contents - Arts & Humanities. Lexikos is also indexed in Web of Science Categories, Language Linguistics and Linguistics.

7.1. Genre Trends of Lexikos

Articles in the journal's own internal system are divided into types such as review articles, contemplative articles, research articles; however, in WoS, by being gathered under a single roof, these types are determined as one type, articles.

Table 1: Types, Numbers and Rates of Records

Document Types	Record Count	% of 441
ARTICLE	336	76.190 %
BOOK REVIEW	60	13.605 %
PROCEEDINGS PAPER	51	11.565 %
EDITORIAL MATERIAL	29	6.576 %
REVIEW	9	2.041 %
BIOGRAPHICAL ITEM	6	1.361 %
NEWS ITEM	1	0.227 %

Table 1 shows the text types and numbers of Lexikos in the WoS Core Collection. Most of the publications of Lexikos in WoS database are articles. 76.190% of the total rate is composed of articles. Although Lexikos includes different genres, a journal mainly deals with lexicography.

Table 2: Lexikos's Number of Publications in WoS by Years

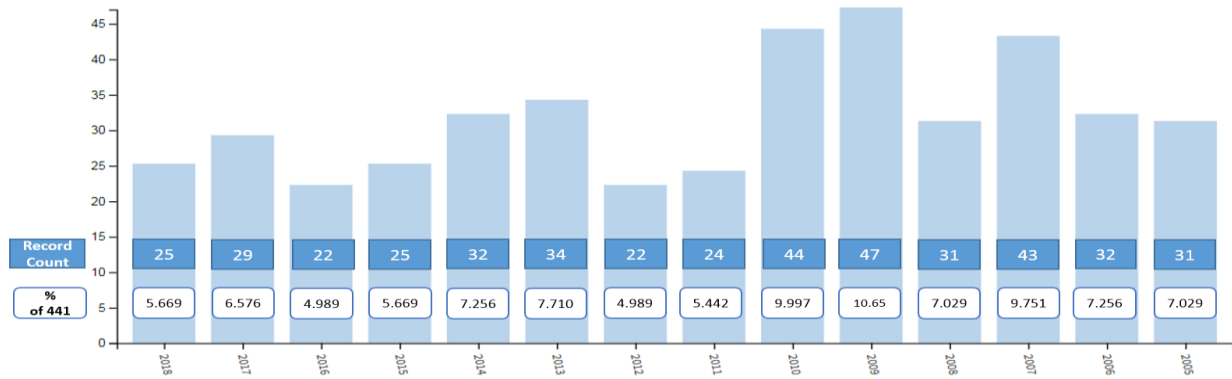
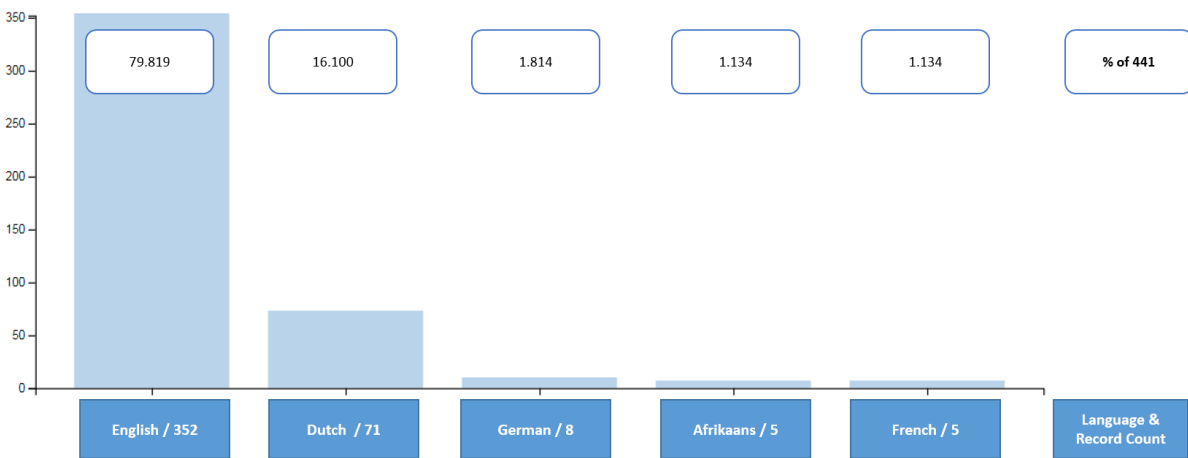


Table 2 shows the numbers and ratios of texts of Lexikos in the Web of Science Core Collection database over the years. The highest record is in 2009. When 2005-2010 and 2011-2018 sections are compared, the number of records in Lexikos in recent years tends to decrease.

Table 3: Numbers and rates of publications in Lexikos according to languages



Publications in this journal can be made in 5 different languages. As can be seen in Table 3, Lexikos's top language for publishing is English. 79.819% of the articles in the journal were published in English. It was published in Afrikaans in 2018, 2017, 2014, 2014, 2005. It was published in French in the following years: 2013, 2009, 2009, 2008, 2005. It has been published in Afrikaans language in recent years, however, the journal has not been published in French for the last 5 years.

7.2. Contributions and Contributors

Between the years 2005-2018, 245 authors contributed to Lexikos Journal without the distinction of the first author, second author, third and fourth author. Some of the texts were produced by single authors and some by several authors; however, in WoS, it is not possible to distinguish between single authors and multiple authors.

Table 4: Number of Authors and Number of Records

Numbers of Authors	Number(s) of Records
149	1
43	2
17	3
9	4
7	5
5	6

149 authors contributed to only 1 (one) publication, 43 authors contributed to 2 (two) publications, 17 authors contributed to 3 (three) publications, 9 authors contributed to 4 (four) publications and 15 authors contributed to 7 or more publications.

7.2.1. The Most Producing Authors

Table 5: The Most Producing Authors

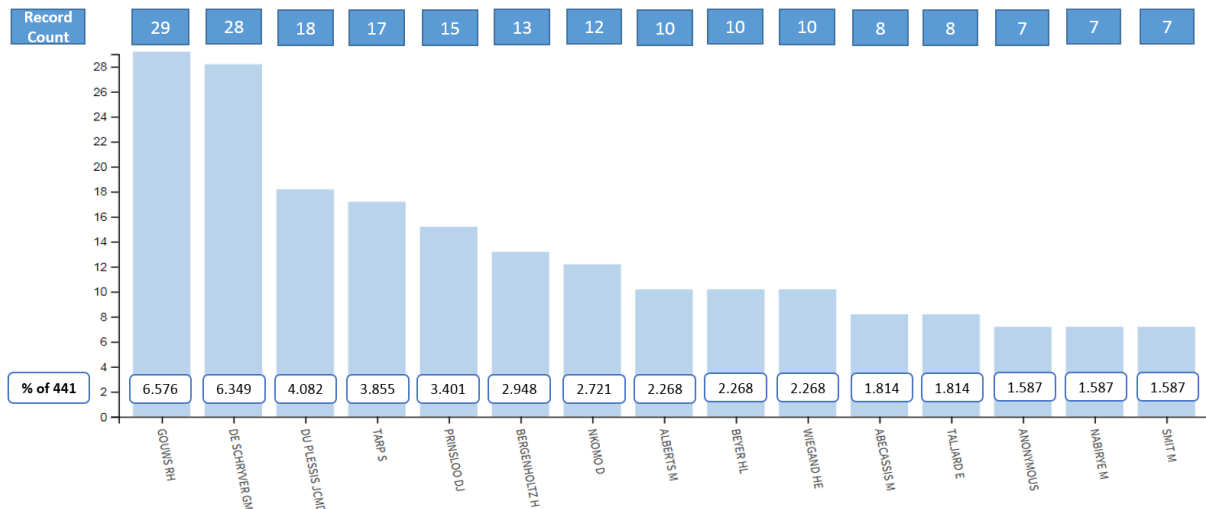


Table 5 shows 15 most producing authors in Lexikos Journal. 7 records are seen in the form of ANONYMOUS because authors' names are not given in Lexicohonour, Lexicotribute, Prepublication Announcements, which are among the types of contributions in the journal. The top 15 authors contributed to the journal published 199 publications that accounted for almost half of the 441 records in the database. Rufus H. Gouws is the author who contributed the most to the journal.

Table 6: The Most Producing Authors and types of their contributions

The Author	ARTICLE	BIOGRAPHICAL ITEM	EDITORIAL MATERIAL	PROCEEDINGS PAPER	REVIEW	BOOK REVIEW
Rufus H. Gouws	23	1	4	2	1	-
de Schryver, Gilles-Maurice	23	-	3	4	1	1
du Plessis, J. C. M. D.	-	-	8	-	-	10
Tarp, Sven	17	-	-	4	-	-
Prinsloo, D. J.	13	-	1	1	-	1
Bergenholtz, Henning	12	-	-	-	1	-
Nkomo, Dion	11	-	-	2	-	1
Alberts, Marietta	8	-	2	-	-	-
Beyer, Herman L.	8	-	2	1	-	-
Wiegand, Herbert Ernst	10	-	-	1	-	-

Table 6 shows which genres in Lexikos are contributed by the most producing authors. These authors were produced texts in the form of articles.

7.2.2. Countries of Authors

Table 7: Countries of Contributors and the numbers of contributions

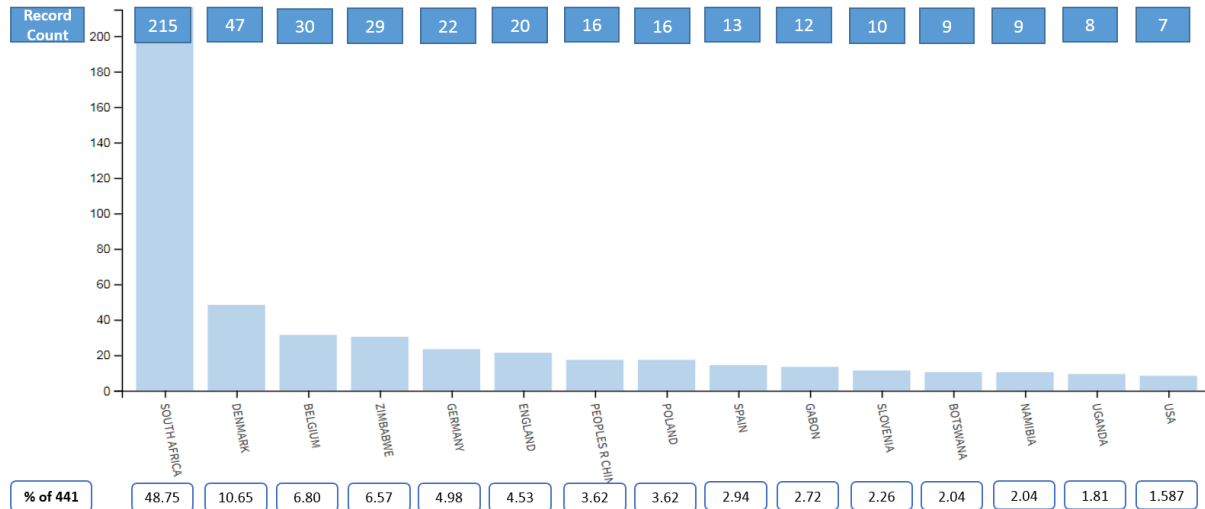
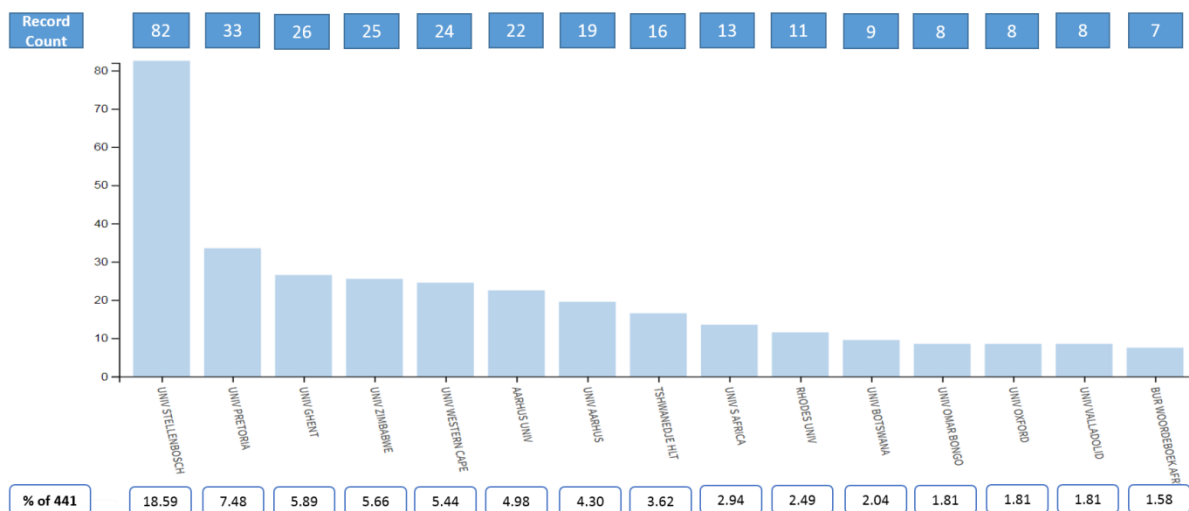


Table 7 shows the countries or regions of the authors who contributed to the Lexikos. Accordingly, the most contributing regions / countries are South Africa and Denmark. These two regions produced more than half of the total records in the journal. Since it is an African based journal, it is normal that the continent of Africa is numerically superior. A total of 39 different countries or regions were contributed to the journal. Only one contribution from 12 countries, two contributions from 8 countries, 3 contributions from 3 countries, 4 contributions from one country were made and the remaining contributions were made from 15 countries. There is no information about 40 records (9.070%) in the database.

7.2.3. Organizations and Affiliations of Authors

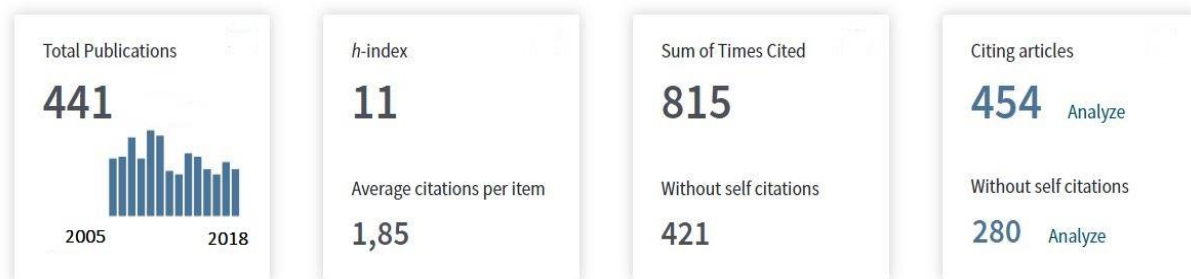
Table 8: Organizations of Contributors



The journal has been contributed by the researchers working in 161 different institutions. There is no institution information about 40 records in the WoS database. Some of the researchers in the field of lexicography sometimes change their institution. Thus, it is possible to face changes in the number of institutions over the years. While ranking the institution in WoS, the institutions of the researchers working in different institutions in the same country are calculated separately; however, their countries are calculated as single. For instance, the country is assumed as one country for an article with three authors from Turkey; and two institutions are counted for three researchers working in two different institutions. The most contributing institution to the journal is Stellenbosch University. 90 institutions in the database contributed with one (1) record, 29 institutions with 2 (two) records and 11 institutions with three (3) records. As in the contributions of the country and the region mentioned in the previous section, the numerical superiority of Africa and Denmark is naturally emerging among the institutions.

7.3. Journal Report of Lexikos

Figure 1: Citation Report of the Lexikos⁴⁵



The Journal has 441 publications in total as Open Access. As can be seen in Figure 1, Lexikos's *h*-index is 11. According to the *h*-index developed by J.E. Hirsch (2005), it is shown that at least 11 publications in the Lexikos received at least 11 citations. When the number of citations (815) to the publications in the journal is divided by the number of publications (441) in the database, the average number becomes 1.85. The sum of times cited in the journal Lexikos is 815. The number of without self citations is 421. The total number of citing articles is 454. The citing articles minus any article that appears in the set of search results (Citing Articles without Self-citations field) is 280. In other words, there are 454 different articles have one citation from Lexikos records. Without self citations means that 280 different articles out of Lexikos cited from Lexikos records.

The number of citations from publications of Lexikos to publications of WoS database (citing article number) is 454. The number of citations in publications in the WoS database but not in Lexikos (without self citations) is 280.

⁴⁵Source:http://apps.webofknowledge.com/summary.do?product=WOS&search_mode=CitationReport&qid=12&SID=C4FEbqy2EAIoIJgyp7&page=1&crNavigationAction=Previous&endYear=44&isCRHidden= Last Accessed March 04, 2019.

Table 9: Sum of Times Cited per Year

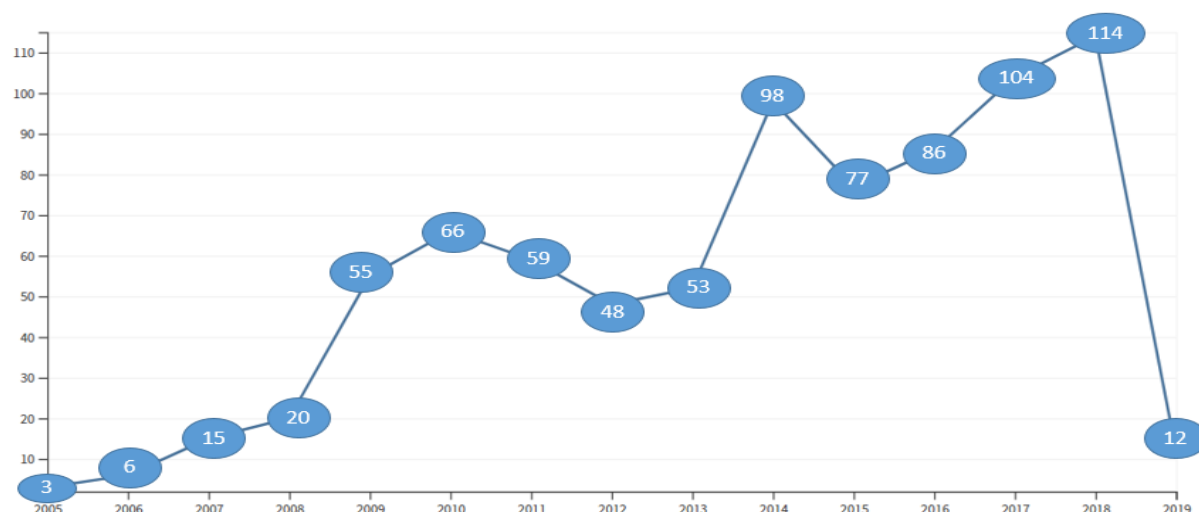


Table 9 shows the number of citations taken by the publications in the journal according to years. According to this, the citation numbers of the articles in the journal increase as we approach the present day. The reason for the decline in 2019 is that this study was carried out in the first months of the year (May, 2019).

Table 10: Most Highly Cited Texts according to WoS Core Collection

	The Name of the text	by	Published year	Times Cited
1	Reflections on Lexicographical User Research	Tarp, Sven	2009	35
2	Foreword	du Plessis, J. C. M. D.	2006	28
3	Do dictionary users really look up frequent words? On the overestimation of the value of corpus-based lexicography	de Schryver, Gilles-Maurice; Joffe, David; Joffe, Pitta; et al.	2006	25
4	The Effect of Lexicographical Information Costs on Dictionary Making and Use	Nielsen, Sandro	2008	21
5	Needs-adapted Data Presentation in e-Information Tools	Bergenholtz, Henning; Bothma, Theo J. D.	2011	18
6	Lexicography in the information age	Tarp, Sven	2007	18

7	Why One and Two Do Not Make Three: Dictionary Form Revisited	Dziemianko, Anna	2012	14
8	How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study	Lew, Robert; Grzelak, Marcin; Leszkowicz, Mateusz	2013	13
9	A Functional Approach to the Choice between Descriptive, Prescriptive and Proscriptive Lexicography	Bergenholtz, Henning; Gouws, Rufus H.	2012	12
10	Multimodal Lexicography: The Representation of Meaning in Electronic Dictionaries	Lew, Robert	2010	12

On the table above, the most cited publications are shown in the WoS database (2005-2018). The most cited publication is “Reflections on Lexicographical User Research” which was prepared by Sven Tarp in 2009.

Table 11: Numbers of citation of Reflections on Lexicographical User Research according to years.

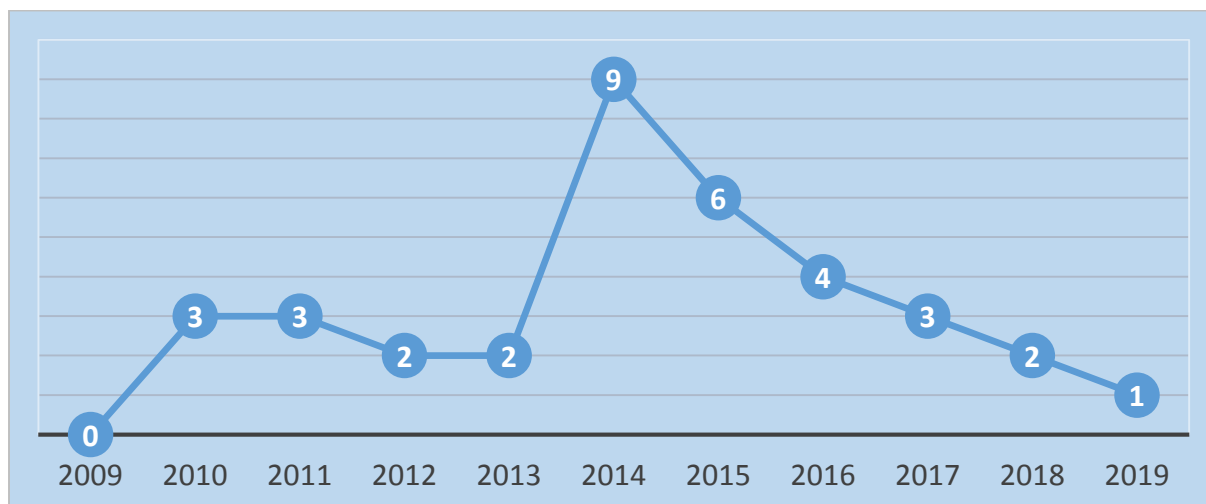


Table 11 shows the number of citations of the article “Reflections on Lexicographical User Research” by years. It is seen that the article peaked in 2014 and then it started to decrease in terms of number of citations after this peak.

Table 12: Most Cited Authors

	Authors	Total Citations
1	Tarp, Sven	108
2	de Schryver, Gilles-Maurice	87
3	Bergenholtz, Henning	73
4	Gouws, R. H.	35
5	Prinsloo, D. J.	33
6	Nielsen, Sandro	31
7	Lew, Robert	30
8	du Plessis, J. C. M. D.	28
9	Beyer, Herman L.	22
10	Wiegand, Herbert Ernst	18

Table 12 shows the most cited authors. Some of these cited studies have multiple authors. When creating this table, the number of citations with multiple-author publications was calculated according to the first author. According to this table, Sven Tarp is the most cited author. The top 10 most cited authors in Lexikos received 465 of 815 total citations.

Conclusion

In this study, it is aimed to describe what is indicated in 14 volumes between 2005 and 2018 in Lexikos by bibliometric method. In order to reveal the scientific impact of the journal, evaluations on its citations have been made.

Of course, the number of citations is not the only important aspect of a scientific publication. However, many countries and institutions attach great importance to the number of citations under the condition of our age. In order to further improve the quality of Lexikos, there are points that should be considered by the journal editorial board and the researchers who will contribute to the journal in the future. First, the number of citations to qualified studies other than the journal should be increased. Publications from outside regions need to be supported in order to spread the journal worldwide. The duty of the researchers is to increase the number of qualified studies that can receive high citation.

Since this study is an oral presentation, the data is limited. Although there are multiple alternatives for citation indexes, Web of Science research system is used. As a future work, thematic trends will be added

to the study using search engines such as <http://www.scielo.org.za> and <https://www.scopus.com> , and an academic article will be created.

References

- Abdi A, Idris N, Alguliyev RM, Aliguliyev RM. (2018). Bibliometric Analysis of IP&M Journal (1980–2015). *Journal of Scientometric Research*. (1):54-62.
- Arik, E. (2015). A bibliometric analysis of linguistics in web of science. *Journal of Scientometric Research*, 4(1), 20-28.
- Day, A, R. (1995). *How to write and publish a scientific paper*, 4Th Edition, Cambridge University Press.
- De Schryver, G. M. (2009a). Bibliometrics in lexicography. *International Journal of Lexicography*, 22(4), 423-465.
- De Schryver, G. M. (2009b). Lexikos at eighteen: an analysis. *Lexikos*, 19 (1).372-403.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46), 16569-16572.
- Lei, L., & Liao, S. (2017). Publications in linguistics journals from Mainland China, Hong Kong, Taiwan, and Macau (2003–2012): A bibliometric analysis. *Journal of Quantitative linguistics*, 24(1), 54–64.
- Lei, L., & Liu, D. (2018). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*.1-23.
- Mohsen, M. A., Fu, H. Z., & Ho, Y. S. (2017). A Bibliometric Analysis of Linguistics Publications in the Web of Science. *Journal of Scientometric Research*, 6(2), 109-18.
- Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25:348-349.
- REF (Research Excellence Framework) (2012). Panel Criteria and Working Methods. https://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf. Last Accessed May 05, 2019.
- Thompson DF. (2018). Bibliometric Analysis of Pharmacology Publications in the United States: A State-Level Evaluation. *Journal of Scientometric Research*,.7(3):167-172.

ON XML-MEDIAWIKI RESOURCES, ENDANGERED LANGUAGES AND TEI COMPATIBILITY, MULTILINGUAL DICTIONARIES FOR ENDANGERED LANGUAGES

Jack Rueter and Mika Hämäläinen

Department of Digital Humanities

University of Helsinki

Abstract

In this paper, we identify the need for a standardized formalism for the structured XML dictionaries of endangered Uralic languages in the Giella infrastructure. For this purpose, we have decided to use TEI formalism as it is a standardized way of representing data and its commonly used in the field of lexicography. This paper focuses on describing the issues and challenges faced in the conversion of the Giella XML into TEI. A full conversion scheme is introduced in this paper contrasting the peculiarities of the two XML formalisms. We incorporate the new TEI-based XML structure into our existing online dictionary system as an output format.

Key Words: endangered languages, XML-MediaWiki, TEI, Uralic languages

Introduction

This paper addresses dictionary-resource development for endangered, under-resourced languages in collaboration with an open-source infrastructure with a rule-based orientation as described in Moshagen et al. (2014). It then outlines advances in XML—MediaWiki synchronization of multilingual dictionaries (Hämäläinen & Rueter, 2018) and enhanced features for etymological and cognate resource work (Hämäläinen & Rueter, 2019) and automatic combination of concepts in multilingual dictionaries (Hämäläinen, Tarvainen & Rueter, 2018). Work with XML introduces a need for a standardized TEI (Text Encoding Initiative) formalism.

As noted in Czaykowska-Higgins (2014), XML structuring greatly benefits from international TEI standards developed since 1990s. Numerous applications bolster personal and professional usage of emerging technologies. Simultaneously, work addresses individual nodes and issues, e.g. etymology (Bowers & Romary, 2016), digitization (Maxwell & Bills, 2017), and endangered language resource development (Czaykowska-Higgins, *ibid.*). The utilization of TEI standard affords shared usage of tools and databases on many platforms as well as multiple possibilities for transformation, rendering and publication.

In the most recent update for TEI (29th January 2019), dictionary guidelines are characterized as catering towards human-oriented presentation. Although readily applicable to majority-language dictionary development, this practice may require tweaking for endangered and low-resourced languages.

Thus, this paper investigates alterations to the orientation in favor of rule-based language-technological infrastructures catering to low-resourced, endangered languages. This entails a strategy of TEI-compatibility, computer-legibility and facilitation of rule-based technologies.

1. TEI-compatibility observed in convertible shorthand XML tags, e.g. $l = \text{lemma}$.
2. Computer-legibility, delimited XML structure depth, unique element-type naming policy.
3. Rule-based description for minimal repetition and expenditures in language-resource development.

Our strategy is to join lexicographical and language-technology efforts for language (re)vitalization, e.g. click-in-text multilingual dictionaries, spellcheckers, etc. This means the introduction of stem-type and inflectional data in lemma-adjacent nodes, something outside the scope of TEI. The solution involves XSLT to formats addressed in Bányi et al. (2017).

Method

A great number of dictionaries in the Giella infrastructure (Moshagen et al., 2014) follow an XML structure that serves a purpose in the infrastructure itself. However, for external use such a format can be seen as troublesome due to insufficient documentation and standardization practices. The fact that these XMLs can be edited in a synchronized way in our MediaWiki environment (Rueter & Hämäläinen, 2017) makes it possible to include new XML formalisms without interfering with the existing Giella infrastructure. As our MediaWiki based online dictionary has been designed with the notion of multiple realizability of the data in different formats, adding a TEI support is just a matter of defining the correspondences between the Giella XML and the TEI standard.

The Giella XML dictionary structure is focused to address issues of machine-readability, minimal weight and reusability. To ensure machine-readability the depth of a given entry element does not exceed four, and element names can only be shared by same-depth elements.

The issue of minimal weight is addressed by establishing mnemonic one-, two- or three-letter element names, which are readily convertible to the TEI standard, but, which for purposes of light infrastructure are used as is in every-day code:

The *e* element stands for entry, this is the base of an entire word article. This element has both attribute and element content. The attribute information address matters of identity (in *id*) and exclusion from specific usage, i.e. exclude generally has the value *fst* (finite-state transducer), which means this particular article is not used in finite-state transducer generation. The minimal contents are one singular *lg* element (the lemma group element) and one or more *mg* elements (meaning group, i.e. sense group). The *lg* element can be preceded optionally by a *map* element, which contains attributes and values pertaining to original dictionary sources, and a *rev-sort_key* element, whose text content consists of the lemma or head word in reverse (right-to-left). The obligatory *lg* element may be immediately followed by a *sources* element with child elements referring to both source literature and parallel attestation of the lemma in other sources. The *resources* element data should, in fact, be directly associated with semantic meaning, and therefore in the future it will be moved to the appropriate *lg* and *mg* subelements.

The *lg* (lemma group) element has no attributes, but it does contain numerous child elements that can be directly associated with word form and not semantics. The two most prevalent child elements of the *lg* are the singular *l* (lemma) and *stg* (stem group) elements. These two elements provide information necessary

for the machine description. Other elements are optional but provide additional information useful in word and word form recognition (*audio*, *etymology*, *compg*, *mini_paradigm*).

The head word text content of the *l* element is augmented by the presence of attributes. These consist, for example, of *pos* (indicating part-of-speech), *hid* (homograph/homonym with values: Hom1, Hom2, ..., whereas the lemma or presentational form of one entry may be identical to that of another, but other morphological forms or origin may distinguish them; only words from the same part-of-speech have distinguishing *hid* attributes), *type* (e.g. common vs. proper noun, where common nouns are default and proper nouns are shown with attributes), *val* (valency of verbs, with initial transitive vs. intransitive marking). All of this information is used in the construction of the finite-state description of a given word in the source language.

The *stg* element has no attributes, and the only child elements it may have are *st* (stem) elements. Each individual *st* element has linguistically relevant text content representing a working morphological stem that all word forms of the paradigm can be derived from in the Giella infrastructure. The attributes, in turn, provide information on inflection type (both for the end user: machine and human), as well as additional data revealing orthographic and language norm status. Since the Giella XML structure allows for pluricentric documentation of a language, i.e. audio and orthographic representations for divergent places in time and space, there are also *varid* (variant identifier) attributes with which to align audio, stem and even possibly *mini_paradigm* content. The *varid* attribute is used in the *st* (stem) element whenever there are more than one *st* element in the *stg* (stem group) element. This serves as a parallel backup to the principle of “prefer first sibling when there are more than one to choose from”, which is necessary when the system is expected to generate a single preferred word form.

The *audio* (this represents audio link information) element is optional. It may have a *varid* (variant identifier) attribute to align it with an *st* sub-element in *stg*. Otherwise it has child elements with information on the audio identifier, the reader, etc. Some of this information could be moved to a different location to minimize the XML content.

In addition to dividing entries according to inflection, they are further divided by an etymological criterion. Thus the *etymology* element only occurs once per entry, and it takes no attributes. It can have multiple etymon and cognate child elements. The etymon element is of mixed content with attributes designating *pos* (part-of-speech), *algu_lekseemi_id* (link information to the external etymological database Álgú), *xml:lang* (639 ISO Language Code reference), in addition to *lemmaID* (lemma identifier) and *stemID* (stem identifier). The cognate element has been used for crosslinking to other dictionaries in the multilingual dictionary set at <https://www.akusanat.com/>.

The *compg* (compound group) element is used for documenting compound words, derivations and inflections alike. While the parent element has attributes *drv* (derivation which can spell out the concatenation with the resulting part-of-speech) and *type* (values are: *Cmp* compound, *Der* derivation, *Infl* inflection), there are ordered *comp* (compound) sub-elements containing link information. The *comp* element has obligatory *ord* (order) attributes to establish constituent order (values: E1, E2,...) although in most instances there are only two *comp* elements. The text content of the individual *comp* element is the lemma or head word, which is then complemented by morpho-syntactic tags in an *msd* attribute, e.g. the value *N.Sg.Gen* might tell us that the specific constituent is a noun appearing in the genitive singular. The *pos* (part-of-speech) attribute here is simply a fallback for when there is no morphological analysis available,

but it is also used to implement crosslinking to the source lemma elsewhere in the dictionary. If the *comp* element contains derivation information, the *pos* attribute value is conceivably *suf* (suffix) and the text content spells out the specific derivation tag used in the Giella infrastructure.

The *mini_paradigm* element provides editors with an opportunity to give feedback on the paradigm produced by the finite-state transducers. These are legacy elements whose output will be used as tickets for prompting improvement in paradigm generation work. In generated pages transducer-produced mini-paradigms will, by default, show the content of the edited mini-paradigm, which can subsequently be turned off by adding an attribute *exclude* with the value *aku* (for the akusanat dictionaries). The *mini_paradigm* element has child elements in *analysis* and grandchild elements in *wordform*. The *analysis* element has an *msd* attribute providing morpho-syntactic analysis with tags separated by full stops. Since there are possibilities of multiple word forms, there may be more than one *wordform* element in which case a *varid* (variant identifier) attribute is necessary.

Results

The corresponding element to the *e* element in TEI is *entry*. As there is no direct correspondence for *lg* in TEI, the information stored in this elements is split into different parts of the TEI structure. The *l* element containing the lemma and part-of-speech is separated into two different tags: *orth* containing the lemma and *pos* under *gramGrp* containing the part-of-speech. The TEI *gramGrp* element also contains the inflectional information from the Giella *stg* and *st* elements under *iType* and *cit*.

The *audio* tags are moved to *cit* elements under *form* element. *Mini_paradigm* is expressed as a *form* element of *infl* type. The *compg* element expressing the compounds that constitute the lemma are split into *cit* elements that project a new dictionary entry structure to express the same information.

The *mg* level is moved to *sense* tags and the *t* elements containing the translations are nested as *cit* elements directly to under the *entry* tag. Finally, the *xg* tags containing the examples are expressed as *cit* elements containing *quote* elements in the TEI structure.

```

1 </e>
2 <lg>
3 <cl pos="N">cuöbbunjuöll</l>
4 <stg>
5 <st ContLex="N_MUORR">cuöb'buöfnjuö%(0%)ll</st>
6 </stg>
7 <audio>
8 <a name="ID_Audio">1129</a>
9 <a name="Reader">23</a>
10 <a name="Recording">1</a>
11 <a name="Included">yes</a>
12 </audio>
13 <mini_paradigm>
14 <analysis ms="Sg.Gen">
15 <wordform-cuöbbunjuöll</wordform>
16 </analysis>
17 <analysis ms="Sg.Ill">
18 <wordform-cuöbbunjuö'lle</wordform>
19 </analysis>
20 <analysis ms="Pl.Gen">
21 <wordform-cuöbbunjuöll</wordform>
22 </analysis>
23 </mini_paradigm>
24 <comp type="Comp">
25 <comp ms="Sg.Gen" ords="E1" pos="N" trans_fin="sammakko">cuöbb</comp>
26 <comp ords="E2" pos="N" trans_fin="nuoili">njuöll</comp>
27 </comp>
28 </lg>
29 <mg relId="0" domain="anatomy">
30 <tg descr_trans="in einem Rentierherz" xml:lang="deu">
31 <t pos="N">Knorpel</t>
32 </tg>
33 <tg descr_trans="in the heart of a reindeer" xml:lang="eng">
34 <t pos="N">cartilage</t>
35 </tg>
36 <tg descr_trans="poron sydämessä" xml:lang="fin">
37 <t pos="N">rusto</t>
38 </tg>
39 <xp>
40 <x src="VJ0:2012:29">cuöbbunjuöllän ceä'lkke</x>
41 <x xml:lang="fin">sammakonnuleksi sanovat.</xt>
42 </xp>
43 </mg>
44 </e>

```

```

1 <entry>
2 <form>
3 <orth-cuöbbunjuöll</orth>
4 <form type="infl">
5 <pron type="Sg.Gen">cuöbbunjuöll</pron>
6 <pron type="Sg.Ill">cuöbbunjuö'lle</pron>
7 <pron type="Pl.Gen">cuöbbunjuöll</pron>
8 </form>
9 <cit type="audio">
10 <cit type="ID_Audio">1129</cit>
11 <cit type="Reader">23</cit>
12 <cit type="Recording">1</cit>
13 <cit type="Included">yes</cit>
14 </cit>
15 <cit type="Cmp" level="0">
16 <cit level="1" type="E1">
17 <quote-cuöbb</quote>
18 <gramGrp>
19 <pos-N</pos>
20 <cit type="nb">Sg</cit>
21 <cit type="case">Gen</cit>
22 </gramGrp>
23 <sense>
24 <cit type="trans" xml:lang="fin">
25 <quote-sammakko</quote>
26 </cit>
27 </sense>
28 </cit>
29 <cit type="E2">
30 <quote-njuöll</quote>
31 <gramGrp>
32 <pos-N</pos>
33 </gramGrp>
34 <sense>
35 <cit type="trans" xml:lang="fin">
36 <quote-nuoili</quote>
37 </cit>
38 </sense>
39 </cit>
40 </form>
41 </entry>
42 <gramGrp>
43 <pos-N</pos>
44 <i type="inflection_type">MUORR</i type>
45 <cit type="inflectional_stem">
46 <quote-cuöb'buöfnjuö%(0%)ll</quote>
47 </cit>
48 </gramGrp>
49 <sense level="0">
50 <sense level="1">
51 <usg type="domain">Anat</usg>
52 <cit type="trans" xml:lang="deu">
53 <quote-Knorpel</quote>
54 <gramGrp>
55 <pos-N</pos>
56 <gen-Msc</gen>
57 </gramGrp>
58 </cit>
59 <cit type="trans" xml:lang="eng">
60 <quote-cartilage</quote>
61 <gramGrp>
62 <pos-N</pos>
63 </gramGrp>
64 </cit>
65 <cit type="trans" xml:lang="fin">
66 <quote-rusto</quote>
67 <pos-N</pos>
68 </cit>
69 <cit type="example">
70 <quote-cuöbbunjuöllän ceä'lkke</quote>
71 <cit type="source">VJ0:2012:29</cit>
72 <cit type="trans" xml:lang="fin">
73 <quote-sammakonnuleksi sanovat.</quote>
74 </cit>
75 </cit>
76 </sense>
77 </sense>
78 </entry>

```

Figure 1: A Giella XML and TEI version of the same entry

Figure 1 shows the structural difference between the existing Giella XML and the TEI XML elaborated in this paper. Both formalisms are capable of representing the same data, but there is a difference in terms of compactness of the two

Discussion

The Giella XML, despite its problems, caters for the need of machine readability and parsability. For this reason the XMLs can be widely used by different tools and services in the infrastructure. The TEI XML introduces more unnecessary complexity for machine readability as the foundations of its design seem to be in preserving the structure of a printed dictionary in a digital format.

However, as for the longevity and reusability of the dictionaries in the future, TEI provides better prospects due to the fact of documentation and standardization. This makes it possible to process TEI XMLs with a multitude of third party applications that provide support for the standard out of the box. Therefore, for us the gain of implementing the TEI formalism is in making the dictionary data available for export in a well-supported format for others to use.

There are a few discrepancies in the XML structure utilized in the Giella infrastructure and those set forth in the TEI standard. While the Giella XML structure caters to dictionary word form generation for multiple reusability, the TEI standard appeals to visual presentation in paper-print and HTML dictionary pages. In practice, the Giella XML has been engineered to serve as a language-independent yet multilingual database, where source- and target-language data are stored in parallel structures, which would allow for language pair flip analysis and sanity checks. The TEI standard offers each individual dictionary project XML structuring possibilities that help guarantee presentation retention for any number of dictionary writing traditions, i.e. the convergence between the Giella XML structure and that of the TEI standard might best be sought in an XSL transformation rendering a bilingual HTML dictionary page.

Conclusions

In this paper, we have presented the existing Giella XML structure used in our MediaWiki based online dictionary. In addition, we have elaborated a way of converting from this XML formalism to the standardized TEI XML. This conversion is provided as an export functionality in our system.

Both the Giella XML and TEI have their own strengths and weaknesses. Supporting both of these formalisms makes it possible for us to combine the best form the both worlds. The Giella XML continues to be the primary import/export formalism for our synchronized MediaWiki-XML dictionary system because of its simplicity and integration with the Giella infrastructure. TEI is introduced as an additional export format for third parties to use the dictionary data in a standardized format.

References

- Bański, P., Bowers, J., & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*.
- Bowers, J. & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. In: *Journal of the Text Encoding Initiative. Issue 10 | 2016 (Open issue) : Selected Papers from the 2015 TEI Conference*.
- Czaykowska-Higgins, E., Holmes, M. D., & Kell, S. M. (2014). Using TEI for an endangered language lexical resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation & Conservation*,
- Hämäläinen, M., & Rueter, J. M. (2018). Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 967-978). Ljubljana: Ljubljana University Press.
- Hämäläinen, M., & Rueter, J. (2019). Finding Sami Cognates with a Character-Based NMT Approach. In *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages: Papers* (Vol. 1, pp. 39-45).
- Hämäläinen, M., Tarvainen, L. L., & Rueter, J. (2018). Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 862-867).

Maxwell, M., & Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 85-91).

Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T., & Tyers, F. M. (2014). Open-source infrastructures for collaborative work on under-resourced languages. *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era* (pp. 71-77).

Rueter, J., & Hämäläinen, M. (2017). Synchronized Mediawiki based analyzer dictionary development. In *3rd International Workshop for Computational Linguistics of Uralic Languages (IWCLUL 2017)* (pp. 1-7).

CORPUS-BASED TERMINOLOGICAL DICTIONARY OF MUSIC: A CASE STUDY OF ROCK GUITAR

Nantakarn Impong and Jirapa Vitayapirak

Faculty of Liberal Arts, King Mongkut's Institute of Technology Ladkrabang

Abstract

This research aims to analyze musical terminology using corpus-based approach. The study started from the survey of user needs analysis for reading materials of rock guitar lessons and then compiling a corpus of rock guitar lessons (RGL). The RGL corpus contains 1,356,029 words (tokens). A concordance program was used to analyze frequency and collocations. The results showed that the vocabularies can be classified into three categories, i.e. general (13.7%), academic (10.5%), and off-list (75.8%). The off-list words were divided into technical terms (8.32%), abbreviations (44.6%), and symbols (22.87%). The results showed that the symbols and abbreviations were frequently used in music lessons. In terms of vocabularies, compounds and multi-words were created from general and academic vocabulary which contains technical meaning of music. These high frequency rock guitar terms, abbreviations and symbols were selected as raw materials for compiling terminological dictionary of Rock Guitar music. In conclusion, this research supports the benefit of the use of specialized corpus to compile terminological dictionaries.

Key Words: Corpus, Rock Guitar, Dictionary, Terminology, KWIC Concordance

Introduction

Rock music is very popular around the world and in Thailand. According to Thai university library statistics survey, there are many guitar references of music written in English. Unfortunately, very few references were written in Thai. There have been no previous studies about the vocabulary levels of rock guitar, especially the technical vocabulary. This research thus focuses on compiling corpus to find out rock guitar terminologies. In other words, this study used corpus-based analysis to study rock guitar lesson texts and lead to design a sample dictionary of rock guitar lessons.

Objectives

1. To analyze technical vocabulary used in rock guitar lessons and examine collocations found in rock guitar lessons.
2. To design a sample dictionary entries based on the technical vocabularies found in the corpus.

Methods

A. Data Collection

Rock guitar lessons were used to compile the Rock Guitar Lesson (RGL) corpus. Three hundred rock guitar lessons from guitar magazines and guitar instruction books were collected as a sample in this study. They were divided into two subcategories, i.e. guitar magazines and guitar instruction books. In

other words, the 150 rock guitar lessons from guitar magazines and 150 rock guitar lessons from guitar instruction books were randomly selected.

B. Research Instruments

In this study selected two instruments: Wordsmith Tools Version 6.0 (Scot 2012), RANGE_GAL_AWL Programs. Wordsmith Tools was used to analyze basic statistics of word analysis in terms of types, tokens, word frequency and concordance. It provided lists of words or word-clusters in texts set out in alphabetical or frequency order, while the concordancer, Concord, offered key-word-in-context (KWIC) concordance displays. Basic statistics such as the number of types, tokens were also examined. The RANGE_GAL_AWL program used was used to calculate the vocabulary levels: General vocabulary, Academic vocabulary, Off-list words refer to Technical vocabulary, Abbreviation, and Symbols.

C. Data Analysis

In this study, the Wordlist, cluster and concordance tools were used to generate lists in frequency order for lexical comparison of texts. In data processing, the frequency and distribution of word types and tokens in the RGL Corpus were first determined. Since the focus of this research was on the terminologies. In order to find out important terms, the three word types, i.e. general (GSL), academic (AWL), and off-list words were analyzed by using RANGE_GAL_AWL Programs. Then, the abbreviations, symbols, and collocation were identified.

Results and Discussion

A .Statistical Analysis of the RGL Corpus

The output of the lexicon extraction program shows the size of the lexicon produced from the RGL Corpus. Statistics can provide various kinds of summary of the contents of the RGL corpus. They can show the Type/Token ratio of the whole vocabulary, which is computed by dividing the number of the tokens by the number of types. It indicates the relative concentration or dispersion of the vocabulary, and offers a measure of its diversity. The lower of type-token ratio, the greater the diversity of words in the corpus.

Table 1: Tokens and Words types in the RGL corpus

Tokens	1,356,029
Word types	40,542
Type/token ratio	1:33.44

In Table 1 the overall corpus size comprised of 1,356,029 tokens or running words. The total number of word types equals 40,542 words in the corpus since a recurrent word is counted only once. For the whole corpus, the ratio of types/tokens is 1:33.44. The ratio indicated that each word was repeated nearly 34 times on average throughout the corpus.

B. Vocabulary Levels

In order to find technical terms in the RGL corpus, the vocabularies were classified into 3 main groups, i.e. General Service List (GSL), academic vocabulary, off-list words consisting of technical, and low frequency words (Nation 2001). It was found out that the three types of vocabulary used in the rock guitar lesson corpus were 13.7% of general vocabulary, 10.5% of academic vocabulary and 75.8% of off-list as shown in Figure1 below:

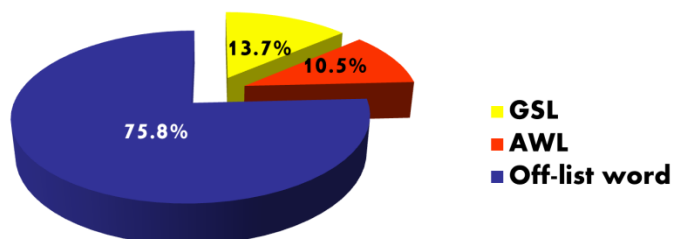


Figure 1: Proportion of General, Academic, and Off-list Vocabularies in RGL Corpus

The off-list words were then divided into 3 groups, i.e. 8.32% of technical vocabulary, 44.6% of abbreviations, and 22.87% of symbols as shown in Figure2 below:

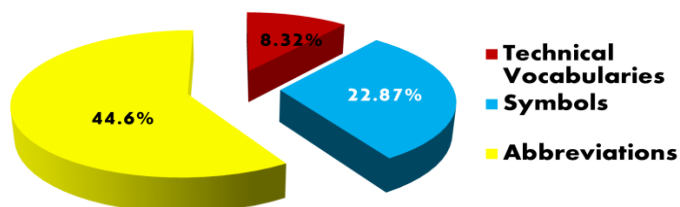


Figure 2: Proportion of Technical, Symbols and Abbreviations in RGL Corpus

This finding did not align itself closely with the figures cited by Coxhead (2000), Nation (2005) suggested that for normal text on average, the GSL and AWL vocabulary represents approximately 90% of running words in academic text. The technical vocabulary covers around 5% and the other low frequency words such as proper name, abbreviation, and numbers represents around 5% of the total words in the text. We can see that the rock guitar lesson is a text-type that covers a large amount of off-word lists, i.e. abbreviations, symbols, and technical terms.

C. Word Classes

In terms of word classes, two major types of word classes, namely function words and content words were found in the corpus. The function words were found at the top ten word frequency lists in the RGL Corpus such as ‘THE’, ‘A’, ‘AND’, ‘TO’, ‘OF’, and ‘IT’.

For a terminological dictionary, content words i.e. high frequency nouns become the focus. Table 3 shows the ten most frequent content words found in the RGL corpus:

Table 2: The Top 10 Content Words in the RGL Corpus

No.	Rank	Words	Freq.	%
1.	19	GUITAR	7,180	0.53
2.	29	CHORD	5,456	0.40
3.	37	MAJOR	4,628	0.34
4.	38	BLUES	4,627	0.34
5.	40	SCALE	4,361	0.32
6.	41	MINOR	4,324	0.32
7.	44	STRING	3,832	0.28
8.	45	CHORDS	3,809	0.28
9.	47	NOTE	3,767	0.28
10.	55	NOTES	3,600	0.27

For Rock Guitar terminological dictionary, the high frequency nouns among content words were concentrated. As shown in Table II, general words of music appeared at the high frequency such as ‘GUITAR’ (7,180 times), ‘CHORD’ (5,456 times), ‘MAJOR’ (4,628 times), ‘BLUES’ (4,627 times), and so on . This finding confirms many studies that scientific and technical discourse makes use of words that are not specialized at all. For example, nouns like ‘guitar, chord, major, blues, scale, minor, string’ are not technical terms. These words are used across many sub-fields. A problem in categorization is the fact that such categories of technical vocabulary in technical texts overlap considerably, due to their use in general language as well as for specific-subject purposes. Examples of this are such as ‘blue, scale, string, chord’ in music. The polysemy of these words in different subject domains underscores the need for an objective means of recognizing the semantic functions of terms in text. In other words, the communicative setting helps to identify meaning and to determine whether an expression will be behaving ‘terminologically’ or ‘normally’.

In fact, technical terms often pair up to form larger meaningful groups within the contexts, i.e. compounds or lexical phrases that always appear in the same form (Hanks, 2010d). For instance, CHORD collocates with PROGRESSION in CHORD PROGRESSION. These collocations can be evidently extracted and identified in context using corpus data in the form of concordances.

D. Compounds

The RGL concordances showed that most of the nouns were compounds consisting of nouns, with preceding nouns functioning as adjectives, e.g. blues guitar, chord progression, bar blues, and so on. We can see that although there are two or more lexemes, the parts function as a single item, with its own meaning. Table 3 below shows the top ten compounds in the rock guitar lesson:

Table 3: The Top 10 Compounds in the RGL Corpus

No.	Word (Freq.)	Collocations	Frequency
1.	GUITAR (7,180)	BLUES GUITAR	1,053
2.	CHORD (5,456)	CHORD PROGRESSION	369
3.	MAJOR (4,628)	MAJOR SCALE	1,317
4.	BLUES (4,627)	BAR BLUES	540
5.	SCALE (4,361)	MAJOR SCALE	904
6.	MINOR (4,324)	MINOR SCALE	449
7.	STRING (3,832)	E STRING	119
8.	CHORDS (3,809)	POWER CHORDS	175
9.	NOTE (3,767)	EIGHT NOTE	114
10.	NOTES (3,600)	QUARTER NOTES	55

E. Technical Vocabularies

In terms of technical vocabularies with high frequency in RGL corpus, the terms such as ‘PENTATONIC’ (1,496 times), ‘ARPEGGIO’ (850 times), and ‘HARMONIC’ (848 times) appeared at the high frequency as shown in Table 4 below:

Table 4: The Top 10 Technical Terms in the RGL Corpus

No.	Rank	Words	Freq.	%
1.	111	PENTATONIC	1,496	0.11
2.	173	ARPEGGIO	850	0.06
3.	174	HARMONIC	848	0.06
4.	210	MELODIC	679	0.05
5.	214	DIMINISHED	664	0.05
6.	271	DOMINANT	511	0.04
7.	318	MODES	437	0.03
8.	332	TAPPING	422	0.03
9.	369	DORIAN	379	0.03
10.	371	CHROMATIC	377	0.03

When observing in-depth about the nature of the specific language of rock guitar lessons using KWIC concordance program, those words do not stand alone. They frequently occurred with other words such as the word ‘PENTATONIC’ always collocates with the word ‘SCALE’ (PENTATONIC SCALE) and used as a compound noun. For the study of collocation, KWIC concordance is a good tool to help finding out the specific collocation of terms. Figure 3 below shows a sample collocation of ‘**melodic minor scale**’:

Figure 3: KWIC Sample of collocation of the word ‘MELODIC MINOR SCALE’

Frequency data of the phraseological terms can be a powerful tool in the hands of technical lexicographers since each level of frequency offers a potential cut-off point for headword selection and

grading of items. The potential headword list for technical dictionary starts with those terms with middle to high frequencies in the corpus. The terms included in the dictionary are examined to make sure that they have technical meanings in the domain of music. Thus, this terminological dictionary is based on corpus frequency information, rather than on introspective intuition or a traditional inventory of words.

Table 5 below shows the high frequency technical terms or multi-word lexical units taken from the top 20 high-frequency content words.

Table 5: Technical Multi-word Lexical Units in the RGL Corpus

No.	Technical Multi-Word Lexical Units	Frequency
1.	MINOR PENTATONIC SCALE	127
2.	A MINOR PENTATONIC	67
3.	MINOR SCALE PATTERN	66
4.	C MAJOR SCALE	65
5.	MELODIC MINOR SCALE	63
6.	HARMONIC MINOR SCALE	60

Regarding collocations, both lexical and grammatical collocations were found in this study. Most technical compounds and multi-word terms consisted of words from GSL, AWL, and technical vocabularies. From Figure 3 and 4 above, it can be seen that the words were not used in general sense but technical musical meaning such as ‘melodic minor scale’. In other words, a close inspection revealed that the collocations in the RGL corpus has a meaning exclusively its own and the meaning cannot be deduced by breaking their parts, e.g. MINOR PENTATONIC SCALE, MELODIC MINOR SCALE. These phrases show highly specific terminological meanings with domain specific implications (Hyland, 2008). The evidence from the corpus suggests which terms are likely to be encountered by language users, and which therefore deserve to be headwords in dictionary.

F. Abbreviations and Symbols

In this study, abbreviations and symbols were frequently found in the corpus due to the fact that the rock guitar lessons include a huge amount of guitar music score. Abbreviation is a short way of writing a word or phrase by leaving out some of letters or by using only the first letter of each word (Crystal, 1992). The top ten abbreviations were shown in Table 6:

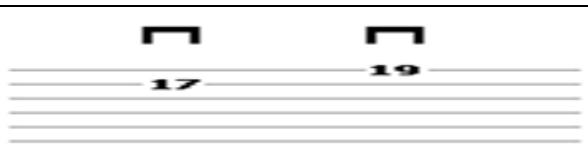
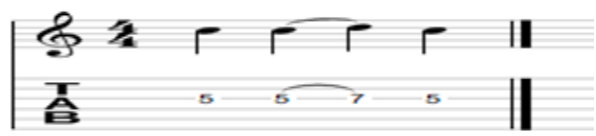
Table 6: The Top 10 High-Frequency Abbreviations in the RGL Corpus

No.	Words	Meaning	Freq.	%
1.	E	E CHORD	11,860	0.87
2.	H	HAMMER ON	10,403	0.77
3.	F	F CHORD	7,751	0.57
4.	G	G CHORD	7,528	0.56
5.	M	MUTING	7,441	0.55
6.	T	TAPPING	7,398	0.55
7.	C	C CHORD	7,046	0.52
8.	S	SLIDE	7,019	0.52
9.	D	D CHORD	5,791	0.43
10.	B	BENDING	5,035	0.37

The high frequency abbreviations include ‘E’ (E Chord-11,860 times), ‘H’ (Hammer on – 10,403 times), and ‘F’ (F Chord – 7,751 times) and so on.

In terms of symbols, the top two high frequency symbols in RGL corpus are shown in Table 7 below:

Table 7: The High-frequency Symbols in RGL Corpus

No.	Symbols	Meaning	Freq.	%
1.		ALTERNATE PICKING	8,579	28.34
2.		HAMMER ON	4,850	16.02

et al. (eds.), Euralex Proceedings. Frisian Institute.

Hyland, Ken. (2008). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*. 27: 4-21.

Scott, M. (2012). *Oxford WordSmith Tools*. Oxford: Oxford University Press.

Nation, P. (2005). Range Program with GSL and AWL. Retrieved from <http://www.victoria.ac.nz/lals/about/staff/paul-nation>

Vitayapirak, J. (2001). *A Corpus-based Approach to ESP Lexicography: The Case Study of English for Thai Computer Science Students*, Macquarie University, PhD. Thesis, Sydney.

COUNT-BASED SEMANTIC MODEL EVALUATION FOR THE EXTRACTION OF SEMANTIC RELATIONS FOR NAMED BAYS FROM A SMALL SPECIALIZED CORPUS

Juan Rojas-Garcia

University of Granada

Riza Batista-Navarro

University of Manchester

Abstract

EcoLexicon (<http://ecolexicon.ugr.es>) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data such as named bays (e.g., *Escambia Bay*, *Narragansett Bay*). For that aim, count-based distributional semantic models (DSMs) were applied to a small-sized, English specialized corpus on Coastal Engineering (7 million tokens) to extract both the terms related to each named bay that is mentioned and the semantic relations. Since the construction of a suitable DSM for a particular task is highly parameterized and its evaluation in small specialized corpora has received little research attention, this paper identified parameter combinations in count-based DSMs appropriate to the extraction of the semantic relations *takes_place_in*, *located_at*, and *attribute_of*, frequently activated by named bays in the corpus. We thus performed an experiment in which count-based DSMs were built on the corpus, and then evaluated on three gold standard datasets for the respective semantic relations, manually extracted from the same corpus and annotated by three terminologists. For our small-sized corpus, the results showed that the log-likelihood association measure outperformed positive pointwise mutual information and *t*-score, and that the detection of a specific relation depended on the context window size.

Key Words: Named bay, Distributional semantic model, Terminology, Knowledge representation, Text mining.

1. INTRODUCTION

EcoLexicon (<http://ecolexicon.ugr.es>) is a multilingual, terminological knowledge base on environmental science that is the practical application of Frame-based Terminology (Faber 2012). The flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas (León-Araúz et al. 2013). However, the representation of geographically contextualized LANDFORM concepts, such as named bays (e.g., *Pensacola Bay*, *Greenwich Bay*), depends on knowing which terms are semantically related to each named landform, and how these terms are related to each other.

With the aim of representing in EcoLexicon the conceptual structures underlying the usage of named landforms in a small-sized, English specialized corpus on Coastal Engineering (7 million tokens), the terms related to each named landform and their semantic relations were extracted with distributional semantic models (DSMs) and other statistical techniques (Rojas-Garcia & Faber 2019). In this paper, we focused on named bays.

DSMs represent the meaning of a term as a vector by considering the statistics of its co-occurrence with other terms in the corpus. Although a DSM can help identify semantic relations between terms based on corpus data (Bernier-Colborne & Drouin 2016), the construction of a suitable DSM for a given task is highly parameterized. Even though numerous studies have addressed the evaluation and optimization of DSMs in very large, general corpora (Baroni et al. 2014; Kiela & Clark 2014; Lapesa et al. 2014), the ability of DSMs to capture different semantic relations in smaller specialized corpora has received little research attention (Fabre et al. 2014).

The objective of this paper was to identify parameter combinations in count-based DSMs for the extraction of three semantic relations, held by named bays, in a small specialized corpus. Hence, the models were evaluated using evaluation data (or gold standards) that contained pairs of semantically related terms, manually extracted from the same corpus. One of the terms was always a named bay, and the other one was an entity or process. The semantic relations that linked the terms were those frequently activated by named bays in the corpus, namely: (a) *takes_place_in* (e.g., STORM SURGE *takes_place_in* ESCAMBIA BAY); (b) *located_at* (e.g., BENTHIC GEOLOGIC HABITAT *located_at* GREENWICH BAY); and (c) *attribute_of* (e.g., WATER QUALITY *attribute_of* NARRAGANSETT BAY). Three gold standard datasets were thus built, one for each of the semantic relations.

The rest of this paper is organized as follows. Section 2 provides background on DSMs, as well as a literature review on their application and evaluation. Section 3 explains the materials, methods, and the DSM evaluation in this study, as well as the construction of the gold standard datasets. Section 4 presents the results obtained. Finally, Section 5 discusses these results along with the conclusions derived and our plans for future research.

2. BACKGROUND AND LITERATURE REVIEW

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller & Charles 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance or cosine similarity (Salton & Lesk 1968), *inter alia*.

Depending on the language model (Baroni et al. 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). The Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde et al. 2006) is an example of this type of model. Prediction-based models exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include continuous bag-of-words (CBOW) and skip-gram models (Mikolov et al. 2013).

Count-based DSMs have been extensively studied (Kiela & Clark 2014; Lapesa et al. 2014; Sahlgren, 2006; Sahlgren & Lenci 2016; Shwartz et al. 2017). Research shows that parameters, such as the context window size, influence the semantic relations that are captured, either syntagmatic relations or paradigmatic relations, namely, synonymy (Curran 2003), antonymy (Santus et al. 2014), hyponymy and meronymy (Shwartz et al. 2017). The syntagmatic relations generally examined are either phrasal associates (e.g., *help* - *wanted*) (Lapesa et al. 2014) or syntagmatic predicate preferences (Erk et al. 2010) in very large, general language corpora. In this study, we focused on the syntagmatic relations

takes_place_in, *located_at*, and *attribute_of*, which were frequently activated by named bays in the specialized language of Coastal Engineering. As far as we know, this framework has not been studied in the context of lexical semantics and DSM evaluation, which constitutes an original aspect of this work.

The capacity of count-based models to detect the previously mentioned syntagmatic relations is described in this paper. Different types of DSM were compared by Baroni et al. (2014), who found that prediction-based models provided better results. In contrast, Ferret (2015) found that count-based models performed better. However, Bernier-Colborne & Drouin (2016) observed that the semantic relations detected by DSMs depended on the window size, and the values for this parameter mostly coincided in both DSMs.

Levy et al. (2015) yielded valuable insights, showing the following: (1) when the parameters of the models were tuned correctly, count-based and prediction-based models obtained similar accuracy; (2) the best model depended on the nature of the task. Nevertheless, Asr et al. (2016) and Sahlgren & Lenci (2016) reported that count-based models outperformed prediction-based models on small-sized corpora of under 10 million words. For that reason, this paper focused on count-based DSMs for a small specialized corpus of 7 million words.

Work in lexical semantics and DSMs includes, *inter alia*, the identification of semantic relations (Bertels & Speelman 2014), information retrieval (Nguyen et al. 2017), word sense discrimination and disambiguation (Pantel & Lin 2002), automatic metaphor identification (Shutova et al. 2010), classification of verbs into semantic groups (Gries & Stefanowitsch 2010), and the use of word vectors as features for automatic recognition of named entities in text corpora (El Bazi & Laachfoubi 2016), and for representation of proper names (Herbelot 2015).

3. MATERIALS AND METHODS

3.1. Materials

3.1.1. Corpus data

The named bays and related terms were extracted from a subcorpus of English texts on Coastal Engineering, on which the DSMs were also built. This subcorpus, comprising roughly 7 million tokens, is composed of specialized texts (i.e., research papers, technical reports, and doctoral thesis) and semi-specialized texts (i.e., handbooks), and is an integral part of the EcoLexicon English Corpus (23.1 million tokens) (see León-Araúz et al. (2018) for a detailed description).

3.1.2. GeoNames geographical database

The automatic detection of the named bays in the corpus was performed with a GeoNames database dump. GeoNames (<http://www.geonames.org>) has over 10 million proper names for 645 different geographical entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored. A daily GeoNames database dump is publicly available as a worldwide text file.

3.1.3. Gold standard datasets

The DSMs, built on our domain-specific corpus, were evaluated on gold standard data, which were manually extracted from the same corpus. The gold standard datasets contained pairs of semantically related terms, in which the semantic relations were those frequently activated by named bays in the

corpus, namely, *takes_place_in*, *located_at*, and *attribute_of*. Three gold standard datasets⁴⁶ were thus built, one for each of the semantic relations. The designations and meaning of these relations are those in EcoLexicon (Faber et al. 2009).

The three semantic relations always linked the normalized designation of a named bay (e.g., *Josiah's Bay* and *Josiah Bay* were normalized to *Josias Bay*) to an entity or process expressed by a nominal term, whether single (e.g., *flooding*) or multiword (e.g., *water quality*). More specifically, the *takes_place_in* relation holds between a process (e.g., *storm surge*, *water level lowering*) and a named bay where the process occurs (see Table 1). The *located_at* relation indicates the location of an entity (e.g., *inundation area*, *shallow estuarine water*) in a named bay (see Table 2). Finally, the *attribute_of* is used for nominal terms that designate characteristics of a named bay (see Table 3).

The three datasets contain 100 instances for each semantic relation, which were all used for the evaluation. The annotation of the pair of terms extracted from the corpus was carried out by three terminologists from our research group. *Cohen's kappa* coefficient was used as the statistical measure of inter-annotation agreement, and the scores for all the annotator pairs were over 90%.

process	<i>takes_place_in</i>	named bay
storm surge	<i>takes_place_in</i>	Escambia Bay
flooding	<i>takes_place_in</i>	Pensacola Bay
geological process	<i>takes_place_in</i>	Narragansett Bay
underwater video imagery	<i>takes_place_in</i>	Greenwich Bay

Examples from the corpus:

1. *Within the Pensacola Bay and Escambia Bay, the shallow estuarine water induces significant **storm surge** at the head of the estuary with a time delay due to the time required for the surge to travel up the estuary.*
2. *Hurricane Ivan causes significant **flooding** over the northeast Gulf coast, especially around **Pensacola Bay** and **Escambia Bay**.*
3. *The objective of this work was to understand the benthic geologic habitats and **geological processes** within two important areas of **Narragansett Bay**.*
4. *Limited **underwater video imagery** was collected in both **Greenwich Bay** (C. Deacutis, personal communication) and Wickford Harbor.*

Table 1: Extract from the gold standard dataset for the *takes_place_in* relation. The last row shows two examples from the corpus.

⁴⁶ The datasets will be available on the website of the LexiCon Research Group at the University of Granada (<http://lexicon.ugr.es>).

entity		located_at	named bay
inundation area		located_at	Pensacola Bay
Blackstone River		located_at	Narragansett Bay
Port Geelong		located_at	Port Phillip Bay
benthic habitat	geologic	located_at	Greenwich Bay
shallow water	estuarine	located_at	Escambia Bay

Examples from the corpus:

1. *Because of the higher simulated storm surge, the **inundation area near Pensacola Bay** from the 3D simulation is slightly more extended than the 2D results.*

2. *Major rivers draining into Narragansett Bay have been extensively dammed, and although not well quantified, models show decreasing sediment load in the **Blackstone River** closer to **Narragansett Bay**, and much of the river is at or close to bedrock.*

3. *The **Port Geelong** located on **Port Phillip Bay** has a significant role in coastal governance arrangements.*

4. *Shumchenia and King (2011) simplified the **benthic geologic habitats** from **Greenwich Bay** into silty (low-energy basin and bay channel habitats) and sandy (depositional platforms and bay-floor sand sheets) geologic habitats.*

5. *Within the Pensacola Bay and **Escambia Bay**, the **shallow estuarine water** induces significant surge at the head of the estuary with a time delay due to the time required for the surge to travel up the estuary.*

Table 2: Extract from the gold standard dataset for the *located_at* relation. The last row shows five examples from the corpus.

characteristic		attribute_of	named bay
water quality		attribute_of	Narragansett Bay
wind speed		attribute_of	Mobile Bay
maximum height	wave	attribute_of	Escambia Bay
high water mark		attribute_of	Pensacola Bay

Examples from the corpus:

1. *The extent of eelgrass (*Zostera marina*), which is an important habitat for juvenile finfish and invertebrates, has declined dramatically in **Narragansett Bay** because of a deterioration in **water quality**.*
2. *Around **Mobile Bay** area, **wind speed** was around 35 m/s or 70 kts.*
3. *Along the I-10 Bridge in **Escambia Bay**, which collapsed during Hurricane Ivan, the simulated **maximum wave height** is about 1.6 - 1.8 m (at about 9 h after the peak wave at Mobile South) with a peak wave period of 8 s.*
4. *Although inadequate data exist for full verification of the maximum inundation map, the simulated and observed **high water marks** at six stations around **Pensacola Bay** and **Escambia Bay** agree well, as shown in Table 3.*

Table 3: Extract from the gold standard dataset for the *attribute_of* relation. The last row shows two examples from the corpus.

3.2. Methodology

3.2.1 Pre-processing

After their compilation and cleaning, the corpus texts were tokenized, tagged with parts of speech, lemmatized, and lower-cased with the Stanford *CoreNLP* package (Manning et al. 2014) for R programming language (R Core Team 2019). The multiword terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

In the DSMs, only terms larger than two characters were considered. Numbers, symbols, and punctuation marks were removed. Since closed-class words are often considered too uninformative to be suitable context words (Kiela & Clark 2014), stopwords were not used (i.e., determiners, conjunctions, relative adverbs, and prepositions). Additionally, the minimal occurrence frequency was set to five so that the co-occurrences were statistically reliable (Evert 2008).

3.2.2. Named bay recognition

Both normalized and alternate names of the bays in GeoNames were searched in the lemmatized corpus. Then, the recognized designations were normalized and automatically joined with underscores. Most bays cited in the corpus were in GeoNames (90%), whereas others were identified by manual inspection (10%). Anaphoric elements referring to a bay were replaced by the corresponding named bay in the

lemmatized corpus. For this task, the automatic anaphora resolution function from *CoreNLP* package was used, and other cases were manually replaced. The 294 bays mentioned in the corpus are shown on the map in Figure 1, with color-coded rectangles that depict their frequency in the corpus. Although latitudes and longitudes could be automatically retrieved from the GeoNames database dump, occasionally, the same designation referred to bays in different countries. For instance, the corpus only located *Botany Bay* in Australia. However, GeoNames indicated that bays with the same name also existed in Canada, the USA, and Virgin Islands. Such cases had to be resolved by corpus queries.

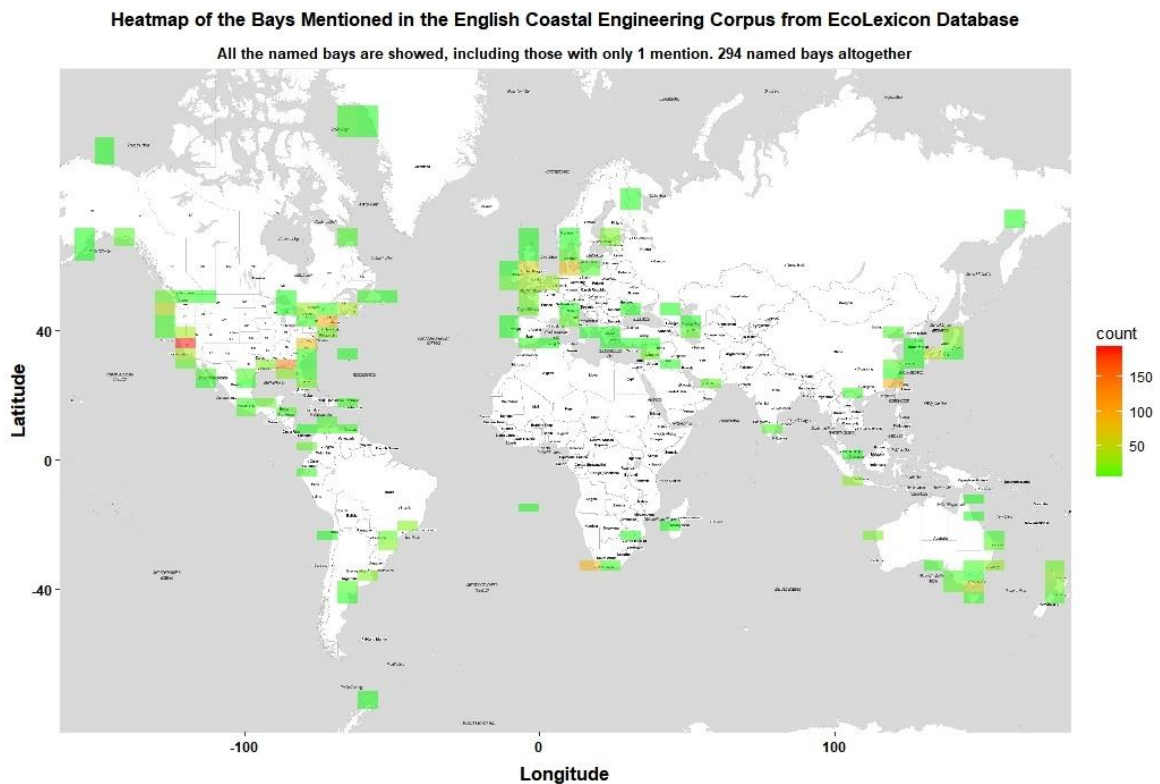


Figure 1: Map with the location and color-coded frequency of the 294 named bays.

3.2.3. Construction of the DSMs

Our experiment involved a comparative evaluation of count-based DSMs for a small-sized specialized corpus, and an exploration of their parameters. The DSMs produced the vector representation of a term based on the contexts in which it appeared in our corpus. In this paper, the contexts of a target term (i.e., a named bay) were the terms that co-occurred with it inside a sliding context window, which spanned a certain number of terms on either side of the target term. These count-based DSMs have various parameters that must be set to build the model, such as the size of the context window. The parameters impinge on both the term representations produced, and the accuracy of the similarity scores between term vectors when the models are compared (Asr et al. 2016; Baroni et al. 2014). Therefore, to assess the ability of the parameters on the DSM to capture the three semantic relations targeted in this paper, various settings for each parameter were essayed, and the combinations of these parameter settings were evaluated. The parameters selected for the count-based DSMs and their settings are described in the following section.

3.2.4. Parameter setting of the count-based models

Count-based models were built with the R package *quanteda* (Benoit et al. 2018) for text mining. To build these models, a term-term matrix of co-occurrence frequencies was first computed, according to a specific size for the sliding context window. Then, the matrix was subjected to a certain weighting scheme, namely, an association measure that increases the importance of the context terms that are more indicative of the meaning of the target term. Subsequently, the association scores were transformed to reduce skewness (Lapesa et al. 2014). A dimensionality reduction technique could also transform this weighted matrix but was not applied. Instead, the 1,000 most frequent words were used, which included all the named bays and terms stored in the three evaluation datasets.

Regarding the context window, we tested size values ranging from 1 to 10 words on either side of the target term, and the context window was allowed to span sentence boundaries. The context window shape was always rectangular (i.e., the increment added to the co-occurrence frequency of a pair of terms was always 1, regardless of the distance between the two terms inside the context window). The frequencies observed on the left and right of a target term were added (this type of window is sometimes referred in the literature to as left+right, or L+R).

Regarding the weighting schemes, three association measures, defined in Evert (2008) were tested: (a) statistical log-likelihood (Dunning 1993); (b) positive pointwise mutual information (PPMI); (c) *t*-score. Log-likelihood and PPMI are widely used in computational linguistics, whereas *t*-score is popular in computational lexicography (Evert et al. 2017).

As reflected in research in computational linguistics, log-likelihood is able to capture syntagmatic and paradigmatic relations (Bernier-Colborne & Drouin 2016: 58; Lapesa et al. 2014: 168), and perform better for medium-to-low-frequency data than other association measures (Alrabia et al. 2014: 4; Krenn 2000). In contrast, PPMI and *t*-score have been found to work well for different applications when compared to other association measures (Baroni et al. 2014; Bullinaria & Levy 2012; Curran 2004; Kiela & Clark 2014). Finally, following Lapesa et al. (2014), the association scores were transformed to reduce skewness. More specifically, log-likelihood and PPMI scores were transformed by adding 1 and then calculating the natural logarithmic (\ln). However, *t*-scores were transformed by calculating the square root ($\sqrt{}$).

The settings tested for each of the two parameters were the following:

1. Size of the context window: 1-10 words (the shape of the context window was rectangular, the type was L+R, and sentence boundaries were spanned).
2. Weighting scheme: $\ln(\log\text{-likelihood} + 1)$, $\ln(\text{PPMI} + 1)$, $\sqrt{t\text{-score}}$.

3.2.5. Evaluation of the count-based DSMs

Once the count-based DSMs were built, they were compared and described, as reported in the following section. First, for each named bay, a sorted list of neighbours was obtained by computing the cosine similarity between the named bay's vector and the vectors of all other context terms. This was also done by sorting then these context terms in descending order of magnitude.

Subsequently, the sorted list of neighbours was evaluated on the whole gold standard dataset for each of the three semantic relations. The measure used to evaluate the models was mean average precision (MAP) (Manning et al. 1998). This measure calculates the accuracy of the sorted list of neighbours obtained for a named bay, based on the rank of its related terms according to the gold standard. The nearer the related terms are to the top of this list on average for each named bay, the higher the MAP.

4. RESULTS

Count-based models were compared by observing the MAP of each model on the three datasets. Table 4 shows the maximum MAP achieved by the models.

Dataset	Count-based models		
	Max. MAP	Weighting scheme	Window size
<i>takes_place_in</i>	0.552 (0.355 ± 0.117)	LL	4
<i>located_at</i>	0.410 (0.302 ± 0.061)	LL	2
<i>attribute_of</i>	0.339 (0.190 ± 0.049)	LL	3

Table 4: Maximum MAP (with average and standard deviation in brackets) of count-based models on each dataset. LL stands for the log-likelihood weighting scheme, transformed by adding 1 and calculating then the natural logarithmic (see Section 3.2.4.).

The results indicated that the *takes_place_in* relation was the most accurately captured by the models, followed by the *located_at* and *attribute_of* relations. The greater accuracy of *takes_place_in* may be due to the large number of instances in specialized Coastal Engineering texts which express the processes that occur in named bays.

As for the *located_at* and *attribute_of* relations, these texts frequently mention the entities in named bays and the characteristics of these landforms. However, it seems that the number of instances of both semantic relations in the whole corpus is not large enough for the DSMs to represent them as accurately as *takes_place_in* instances.

Table 4 shows that the maximum MAP of count-based models were achieved when:

1. The statistical association measure for the three semantic relations was log-likelihood, transformed by adding 1 and calculating the natural logarithmic (indicated with the abbreviation LL in Table 4).
2. The window size for the *takes_place_in* relation was 4 words.
3. The window size for the *attribute_of* relation was 3 words.
4. The window size for the *located_at* relation was 2 words.

To assess the impact of the window size on the accuracy of the count-based models, the average MAP for each setting of this parameter (i.e., for each window size between 1 and 10 words) is shown in Figure 2. The average MAP was used, instead of the maximum, because it allowed us to determine which window size settings consistently produced satisfactory results, regardless of the settings used for the weighting scheme parameter.

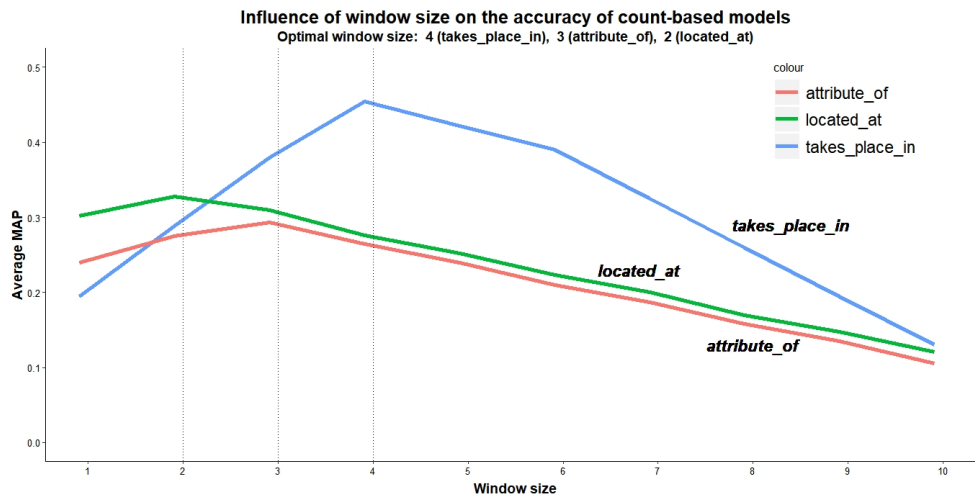


Figure 2: Average MAP of count-based models with regard to window size.

Figure 2 indicates that the optimal size was 4 words for the *takes_place_in* relation, 3 words for *attribute_of*, and 2 words for *located_at*.

According to the results for the window size, the processes that occur in named bays tend to appear in sentences at an average distance of 4 words from the named bay, without counting closed-class words such as prepositions (e.g., *inside*, *within*, *of*), conjunctions (e.g., *because*, *and*), relative adverbs (e.g., *where*), and determiners (e.g., *the*, *a*), which were removed in the pre-processing stage. This can be verified in example (1) from the corpus, which contains the *takes_place_in* relation to *Escambia Bay* and the process *storm surge*, which appears at a distance of 4 words (underlined in the sentence) on the right of the named bay (multi-word terms such as *shallow estuarine water*, joined with underscores in the lemmatized corpus, were considered one word):

- (1) *Within the Pensacola Bay and **Escambia Bay**, the shallow estuarine water induces significant storm surge at the head of the estuary with a time delay due to the time required for the surge to travel up the estuary.*

Obviously, the processes that occur in named bays are also situated at different distances. In example (2) from the corpus, the process *hurricane wind* is 2 words on the right:

- (2) *On the western side of the eye, inside **Mobile Bay** and Mississippi Sound, **hurricane wind** (directed offshore) is much weaker partly due to the dissipation of the open land.*

Regarding the *attribute_of* relation, although 3 words was the optimal window size for an accurate representation in the DSMs, as shown in example (3) (*Escambia Bay* is related to *maximum wave height*), the nominal terms that characterize named bays are also situated at 2 words from them, as in example (4):

- (3) The simulated maximum wave height and wave direction inside the Pensacola Bay and Escambia Bay is shown in Figure 10b.
- (4) The extent of eelgrass (*Zostera marina*), which is an important habitat for juvenile finfish and invertebrates, has declined dramatically in Narragansett Bay because of a deterioration in water quality.

The *located_at* relation required an optimal window size of 2 words, a situation illustrated in example (5) from the corpus, but the entities placed in named bays are also at a distance of only 1 word, as in example (6):

(5) The Port Geelong located on Port Phillip Bay has a significant role in coastal governance arrangements.

(6) The bars in Greenwich Bay likely form and are only modified during storm events.

5. CONCLUSIONS

To extract knowledge for the representation in EcoLexicon of the conceptual structures that underlie the usage of named bays in a small Coastal Engineering corpus (Faber 2012), count-based DSMs were applied to the corpus to extract the terms related to each named bay. Since the construction of DSMs is highly parameterized, and their evaluation in small specialized corpora has received little research attention (Fabre et al. 2014), this paper identified parameter combinations in count-based models suitable for the extraction of the semantic relations *takes_place_in*, *located_at*, and *attribute_of*, frequently held by named bays in the corpus. The models were thus evaluated using three gold standard datasets.

The log-likelihood association measure showed the best performance for the three semantic relations. This result reinforces the findings of previous research, which states that log-likelihood achieves greater accuracy for medium-to-low-frequency data than other association measures (Alrabia et al. 2014: 4; Krenn 2000). The optimal window size depended on the semantic relation that was to be captured, namely, a window size of 4 words for the *takes_place_in* relation, 3 words for *attribute_of*, and 2 words for *located_at*. The dependence of the window size on the specific semantic relation is in line with previous results reported in the literature (Bernier-Colborne & Drouin 2016: 57; Lapesa et al. 2014).

It was also found that the *takes_place_in* relation was the most accurately represented by the count-based models, followed by *located_at* and *attribute_of*. This was possibly due to the insufficient number of instances of both semantic relations in the corpus for the DSMs to represent them as accurately as *takes_place_in* instances.

Future work will include testing the following: (a) prediction-based DSMs, such as GloVe (Pennington et al. 2014), CBOW and skip-gram (Mikolov et al. 2013), and fasttext (Bojanowski et al. 2017); (b) other parameters, such as the shape of the context window; (c) the application of the dimensionality reduction technique for texts Topic Modeling (Blei et al. 2003). Furthermore, the DSMs will be evaluated on gold standard datasets for named rivers and beaches, and the three datasets for named bays, used in this study, will be increased with more annotated data.

Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author.

References

- Alrabia, M., Alhelewh, N., Al-Salman, A. & Atwell, E. (2014). An empirical study on the Holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics* 5(1), 1-13.
- Asr, F., Willits, J. & Jones, M. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In A. Papafragou, D. Grodner, D. Mirman & J. Trueswell (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci)*, Philadelphia (Pennsylvania), 10-13 August, 2016, 1092-1097.
- Baroni, M., Dinu, G. & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, Baltimore, 22-27 June, 2014, 238-247.
- Benoit K, Watanabe K., Wang H., Nulty P., Obeng A., Müller S., & Matsuo A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774.
- Bernier-Colborne, G. & Drouin, P. (2016). Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm)*, Osaka, 12 December, 2016, 52-61.
- Bertels, A. & Speelman, D. (2014). Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology* 20(2), 279-303.
- Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993-1022.
- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, 135-146.
- Bullinaria, J.A. & Levy, J.P. (2012). Extracting semantic representations from word cooccurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods* 44, 890-907.
- Curran, J.R. (2003). *From distributional to semantic similarity*. Doctoral dissertation, University of Edinburgh.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61-74.
- El Bazi1, I. & Laachfoubi, N. (2016). Arabic named entity recognition using word representations. *International Journal of Computer Science and Information Security* 14(8), 956-965.
- Erk, K., Padó, S. & Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics* 36(4), 723-763.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter, 1212-1248.
- Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-alation – A large-scale evaluation study of association measures for collocation identification. In *Proceedings of the eLex 2017 Conference*, Leiden, 19-21 September, 2017, 531-549.
- Faber, P. & Mairal, R. (2009). *Constructing a Lexicon of English Verbs*. Berlin & New York: Mouton de Gruyter.

- Faber, P. (ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.
- Faber, P., León-Araúz, P. & Prieto, J.A. (2009). Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies* 1, 1-23.
- Fabre, C., Hathout, N., Sajous, F. & Tanguy, L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *TALN-RECITAL 2014 Workshop SemDis 2014 : Enjeux actuels de la sémantique distributionnelle (SemDis 2014: Current Challenges in Distributional Semantics)*, Marseille, July, 2014, 266-279.
- Ferret, O. (2015). Réordonner des thésaurus distributionnels en combinant différents critères. *TAL* 56(2), 21-49.
- Gries, S. & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice, S. & J. Newman (eds.) *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford (California): CSLI, 73-90.
- Herbelot, A. (2015). Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, London, 15-17 April, 2015, 151-161.
- Kiela, D. & Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, 26-30 April, 2014, 21-30.
- Krenn, B. (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. Saarbrücken: DFKI & Universität des Saarlandes, vol. 7, Saarbrücken Dissertations in Computational Linguistics and Language Technology.
- Lapesa, G., Evert, S. & Schulte im Walde, S. (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, Dublin, 23-24 August, 2014, 160-170.
- León-Araúz, P., Reimerink, A. & Faber, P. (2013). Multidimensional and multimodal information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem & P. Fuglewicz (eds.) *Computational Linguistics*. Berlin: Springer, 143-161.
- León-Araúz, P., San Martín, A. & Reimerink, A. (2018). The EcoLexicon English corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress*, Ljubljana, 17-21 July, 2018, 893-901.
- Levy, O., Goldberg, Y. & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3, 211-225.
- Manning, C.D., Raghavan, P. & Schütze, H. (1998). *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.
- Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore (Maryland), 23-24 June, 2014, 55-60.

- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*, Scottsdale, 2-4 May, 2013.
- Miller, G.A. & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1-28.
- Nguyen, N.T.H., Soto, A.J., Kontonatsios, G., Batista-Navarro, R. & Ananiadou, S. (2017). Constructing a biodiversity terminological inventory. *PLoS ONE* 12(4), e0175277.
- Pantel, P. & Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, Edmonton, 23-26 July, 2002, 613-619.
- Pennington, J., Socher, R. & Manning, C.D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing (EMNLP)*, Doha (Qatar), 25-29 October, 2014, 1532-1543.
- R Core Team (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Reimerink, A. & León-Araúz, P. (2017). Predicate-argument analysis to build a phraseology module and to increase conceptual relation expressiveness. In R. Mitkov (ed.) *Computational and Corpus-Based Phraseology*. Cham (Switzerland): Springer, 176-190.
- Rohde, D., Gonnerman, L. & Plaut, D. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM* 8, 627-633.
- Rojas-Garcia J. & Faber, P. (2019). Extraction of terms for the construction of semantic frames for named bays. *Argentinian Journal of Applied Linguistics* 7(1): 27-57.
- Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral dissertation, Stockholm University.
- Sahlgren, M. & Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (Texas), 1-5 November, 2016, 975-980.
- Salton, G. & Lesk, M.E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM* 15(1), 8-36.
- Santus, E., Lu, Q., Lenci, A. & Huang, C-R. (2014). Unsupervised antonym-synonym discrimination in vector space. In R. Basili, A. Lenci & B. Magnini (eds.) *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, Pisa, 9-11 December, 2014, vol.1, 328-333.
- Schmidt, B. & Li, J. (2017). *wordVectors. Tools for creating and analyzing vector-space models of texts*. R package version 2.0.
- Shutova, E., Sun, L. & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, 23-27 August, 2010, vol. 2, 1002-1010.
- Shwartz, V., Santus, E. & Schlechtweg, D. (2017). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, Valencia, 3-7 April, 2017, vol. 1, 65-75.

DICTIONARY OF OLD RUSSIAN PLANT NAMES (11TH–17TH CC.): WORD ENTRIES DRAFTS

Kira I. Kovalenko

Valeria B. Kolosova

Institute for Linguistic Studies, Russian Academy of Sciences

Abstract

In the article, the structure of the prospective Old Russian plant names dictionary (the 11th–17th cc.) is discussed. Being created on the basis of the PhytoLex Russian plant names database, it will include the majority of the data represented there. The dictionary is going to be compiled in Lexonomy using XML for the data modeling, which means that the first issue for the compilers is to develop an appropriate XML model, following TEI recommendations. The first part of the dictionary will represent Old Russian words and their meanings and will contain such types of information as forms, gender, phonological and orthographic variants, etymology, definition(s), illustrations with quotations and information about their sources. Additional materials may contain geodata (places of growth and trade), functions of the plant (medicinal, magical, religious, etc.), contextual usage (e.g. metaphorical), synonyms, derivatives, and pictures (if available). The second part will give the opportunity to find all the names of a particular plant recorded in the Russian language from the 11th to the 17th century, which can be quite numerous. The electronic format of the dictionary allows several people to work simultaneously, which is very important in a joint project. It also helps with representing new data in real time and easily preparing printed versions if needed. The dictionary is intended to be an open resource, useful for both ethnobotanists and specialists in the Slavic historical lexicology. It will also become the basis for research into the later periods of the Russian language.

Key Words: PhytoLex, Old Russian language, phytonyms, electronic dictionaries, Lexonomy

Introduction

Old Russian plant names is one of the most difficult lexico-semantic groups to study. Word extraction and analysis are extremely time consuming, which causes lacunas and mistakes even in academic historical dictionaries, i.e. [Dictionary 11th–14th; Dictionary 11th–17th; Dictionary 16th–17th]. The newly created linguistic database PhytoLex (<https://phytonyms.iling.spb.ru>) is designed for registering plant names in the Russian text from the 11th up to the 17th centuries as completely as possible. The basic purposes are: to collect plant names mentioned in Russian texts from the 11th to the 17th centuries; supply them with contexts describing their outlook, functions and legends connected with them; identify unknown Russian plant names; find out the time and mechanisms of borrowing (for loanwords), changes in their spelling and meaning (see details in [Kolosova et al. 2018]).

Methods

The materials for the database are taken by the full-view method from historical dictionaries, archival materials, published and unpublished manuscripts, dictionaries, compiled by foreigners working in Russia, travelogues, medical documents, etc. It is being constantly replenished by new plant names,

added by the project members, each of whom is responsible for certain genres: G. Molkov (Old Russian texts of the 11th-15th cc.), A. Ippolitova (herbal books), O. Olekhovich, K. Khudin (documents of the Apothecary Chancery), K. Kovalenko (handwritten lexicons), A. Shchekin (historical documents), V. Kolosova (foreign dictionaries of the Russian language), K. Zaytseva (technical implementation). The PhytoLex database was created in the Django framework; for the Web interface the Bootstrap framework was used. The data are stored in an open-source object-relational database system, PostgreSQL.

The word entries were compiled in Lexonomy (www.lexonomy.eu) — a professional platform for writing and publishing dictionaries (see the description in [Měchura 2017]). The possibility of its use by multiple compilers makes it possible to work collectively on the dictionary. As the preliminary materials are represented in the database, including identification of the plant and its functions, the process of compilation of word entries includes unification of the headword variant forms (which are given separately in the database) and picking up the most representative citations.

Results

Along with the plant names and their etymology, the database contains a lot of related information: authorship of texts (if known), place and time of creating the document, functions of the plant, etc. Currently (May 2019), the PhytoLex contains 511 plants, 249 sources, 56 publications, 416 etymologies, and 1,797 lexemes, which are reflected in 5,460 word usages. Search options allow looking for such information as names for the same plant; plants having the same name; phytonyms from the same source; phytonyms from the same time period, etc. All this may serve as a sufficient basis for the future Russian plant names dictionary of the 11th-17th C. compilation of which is to be started after the database has representative volume. Nevertheless, word entry structure is to be developed and some word entries can be represented for discussion in advance to foresee all the details and avoid possible mistakes.

Discussion

The dictionary word entries contain the same types of information as the database does: headword, phonological and orthographic variants, etymology, meaning(s), illustrations with quotations and information about their sources, geodata (places of growth and trade), functions of the plant (medicinal, magical, religious, etc.), contextual usage (e.g. metaphorical), synonyms, derivatives, and pictures (if available).

The schema of the dictionary is based on the TEI P5 recommendation (www.tei-c.org), although some elements had to be included additionally. They are: <function> and <derivative> containing a description of the plant's cultural function and derivative forms.

The simple word entry, containing the headword, grammatical information, etymology (if known), definition, citation(s), and bibliographical data can be represented as follows:

```
<entry>
  <form>
    <orth>каморнумъ [kamornum]</orth>
    <gram>м. [m]</gram>
  </form>
```

```

<etym>от <lang>греч. [Greek]</lang> <foreign>συκόμορος</foreign>
</etym>
<sense>
<def>
<lat>Ficus sycomorus</lat>, сикомор.</def>
<cit>
<quote>Икамориумъ, или каморнумъ, ягодичина. [Ikamorium, or katornum, sycamore fig.]</quote>
<bibl>
<title>Азбуковник Давида Замарая [Azbukovnik of David Zamaray]</title>
<date when='1626'>1626</date>
</bibl>
</cit>
</sense>
</entry>

```

The `<lat>` tag helps to index Latin plant names, which ideally are identified up to the species and are given according to the standardised classification represented in the *Catalogue of Life* portal database [Roskov et al. 2019]. Accurate identification can be very difficult; for example, the plant now known as *Matricaria chamomilla* L. had as many as 36 synonyms at different times and in different catalogues (<http://www.catalogueoflife.org/col/details/species/id/28ea026a937c582fade5a9f0cbd4a20b>), and Old Russian names could be supplied with any of them, if it had an identification at all.

Nevertheless, only one Latin name marked in the database as *accepted name* should be given in a definition — which gives us the possibility to search for all the Old Russian names of a particular plant, which also can be quite numerous.

Unfortunately, that is not always possible, as sometimes there the genus only is known, or there is no evidence mentioned to give any identification at all. For example, the following entry contains only the genus in its definition:

```

<entry>
<form>
<orth>Гуль [Gul']</orth>
<gram>(?) [gender is not defined]</gram>
</form>
<etym>от <lang>тур. [Turkish]</lang> <foreign>gül</foreign>
</etym>
<sense>
<def>

```

<lat>Rosa</lat> роза [rose].</def>

<cit>

<quote>Трандафило, родонъ есть цвѣтъ от плода терновнаго, по словенски свороборинной цвѣтъ, благоуханен, и многолиствен и красен, а по латыне роса; есть древо триандафило, величиною съ человека и ниже, цвѣтъ на нем красной и бѣлой, что маковой, а пахнетъ благоуханно, а по турски гуль, а цвѣтъ ставятъ в печи знойные на ночь, и ту воду пьютъ, то есть вода благоуханнаа. [Trandafilo, rodon is a flower of the fruit of blackthorn, in Slavic *svoroborina* flower, fragrant, and multi-petaled and beautiful, and in Latin called *rosa*; there is a *triandafilo* tree, having the size of a man and lower, the flowers on it are red and white, like poppy's, and it smells fragrant, and in Turkish is called *gul*, and the flowers are put in hot ovens for a night, and that water is for drinking, that water is fragrant.] </quote>

<bibl>

<title>Азбуковник Сергия Шелонина [*Azbukovnik* of Sergy Shelonin]</title>

<date notBefore='1653' notAfter='1659'>1653-1659</date>

</bibl>

</cit>

</sense>

</entry>

In the case where the headword has several phonetic, orthographic, or morphological variants, they will be given there as well. Most often they represent the process of adaptation of the foreign word in the Russian language. Variant forms are shown as follows:

<entry>

<form>

<orth>агримония [*agrimoniya*]</orth>

<gram>ж. [f.]</gram>

</form>

<form>

<orth>агримониумъ [*agrimonium*]</orth>

<gram>неизм. [unchangable]</gram>

</form>

<etym>от <lang>lat.</lang> <foreign>agrimonia</foreign>

</etym>

<sense>

<def>

<lat>Agrimonia eupatoria L.</lat>, репешок обыкновенный [common agrimony].</def>

<cit>

<quote>Трава агримонии. [Herb agrimonii.]</quote>

<bibl>

<title>Рецепты Аптекарского приказа [Prescriptions of the Apothecary Chancery]</title>

<num>143-2-1059</num>

<date when='1673'>1673</date>

</bibl>

</cit>

<cit>

<quote>Травы агримониумъ [of herb agrimonium].</quote>

<bibl>

<title>Рецепты для боярина Ильи Даниловича Милославского [Prescriptions for Иуа Danilovich Miloslavsky]</title>

<num>143-2-743</num>

<date notBefore='1665' notAfter='1666'>1665-1666</date>

</bibl>

</cit>

</sense>

</entry>

<entry>

<form>

<orth>Шаптала [*shaptala*]</orth>

</form>

<form>

<orth>шептола [*sheptola*]</orth>

</form>

<GramGr>

<gen>ж. [f.]</gen>

</GramGr>

<etym>от <lang> тур. [Turkish]</lang> <foreign>šäftaly</foreign> < <lang> перс. [Persian] </lang> <foreign>šäftälū</foreign>

</etym>

<sense>

<def>

<lat>Prunus persica (L.) Stokes</lat>, персик [peach].</def>

<cit>

<quote>А овощей всяких много, яблок, и груш, и вишен, и слив, и дынь и арбузов, и винограду, и огурцов, и орехов гредцких и русских, и меду, и сахару леденцу и белого, и шапталы, а иные овощи и неведомы. [And vegetables are numerous, apples, and pears, and cherries, and plums, and melons and watermelons, and grapes, and cucumbers, and nuts from Greece and Russia, and honey, and sugar — both lollipop sugar and the white one, and *shaptalas*, and some vegetables are unknown.]</quote>

<bibl>

<title>Первые русские дипломаты в Китае, [The first Russian diplomats in China]</title>

<num>125</num>

<date when='1654'>1654</date>

</bibl>

</cit>

<cit>

<quote>А яблоки и груши, и сливы, и дыни, и шептолы поспевают на Петровъ день и ранее. [And apples and pears, and plums, and melons, and *sheptolas* ripen on St. Peter's day and earlier.]</quote>

<bibl>

<title>Первые русские дипломаты в Китае [The first Russian diplomats in China]</title>

<num>125</num>

<date when='1654'>1654</date>

</bibl>

</cit>

</sense>

</entry>

As it is common in the [Dictionary 11th–17th], it is proposed to supply each citation with the time of the copy creation and the time of the supposed creation of the text. For example, citations from the *Life and Pilgrimage of Danylo, Hegumen from the Land of the Rus* for the database were taken from the published copy of 1495 while the *Life and Pilgrimage* was supposedly created between 1104 and 1106 [Prokhorov 1997]. Both dates should be given after the citation:

<entry>

<form>

<orth>верба [*verba*]</orth>

</form>

<GramGr>

<gen>ж. [f.]</gen>

</GramGr>

<sense>

<def>

<lat>Salix sp. L.</lat>, ива [willow].</def>

<cit>

<quote>Есть же по сей странѣ Иордана на купѣли той яко лѣси древо не высоко, аки вербѣ подобно есть, и выше купѣли тоя по берегу Иорданову стоитъ яко лозие много, но нѣсть якоже наша лоза, но нѣкако аки силяжи подобно есть; есть же и тростие много; болоние имать, яко Сновь рѣка. [On this bank of the Jordan, near that font is a kind of a wood; the trees are low, similar to willows, and higher up the river from that font, along the banks of the Jordan, kind of osier are standing in great number, but it is not the same as our osier, it looks like a cornel a little; there is also a lot of reeds, there are backwaters, like on the river Snovj.]</quote>

<bibl>

<title>Хождение игумена Даниила [The Pilgrimage of Abbot Daniel]</title>

<date when='1495'>1495</date>

<origDate notBefore='1104' notAfter='1106'>1104-1106</origDate>

</bibl>

</cit>

</sense>

</entry>

The important part of the plant representation is information on the plant parts and the plant functions in culture and everyday life. The following plant functions are given in the list in the database:

- veterinary medicine (worms, cough, diarrhea, prophylaxis; rabbits, poultry (geese, goslings, turkeys, hens, chickens, swans, ducks, ducklings), cattle (goats, cows, horses, sheeps, pigs, calves);
- decorative (bouquet, wreath, wedding wreath);
- cosmetics (white, eyebrow dye, hair dye, lipstick, blush, mascara);
- dyeing (yellow, green, brown, red, blue, black);
- magic (apotropaic, military, imitative, carpogonic, social, hunting, love spells, agricultural, etymological);
- material (barrels, ropes, hedges, icons, baskets, roof, furniture, brooms, floors, dishes, piles, building, church plates);
- medicine (whites, insomnia, hypertonia, hypotonia, worms, deafness, sore throat, hernia, baby crying, diabetes, constipation, cough, bloody diarrhea, thrush, runny nose, dumbness, frostbite,

burns, poisoning, diarrhea, cuts, cold, bruises, blindness, sunstroke, abrasions, heatstroke, bruises, ulcer);

— honey plant;

— drinks (cocoa, jelly, coffee, mead, tincture, beer, *sbiten*, alcohol, tea).

— environment (shadow);

— food (jam, side dish, porridge, marinades, chowder, seasoning, salads, bakery, pickles, soup, bread, spices);

— religion (feeding ancestors, May tree, ritual symbol, donation, wedding tree, altar decoration, home decoration, icons decoration, decoration of participants of calendar rites, church decoration, covering the road to the cemetery, covering floor);

— social (smoking, snuffing tobacco);

— fodder (rabbits, poultry (geese, goslings, turkeys, hens, swans, ducks, ducklings, chickens), cattle (goats, cows, horses, sheep, pigs, calves));

— economic (aromatization, tar, lighting, writing materials, fuel, coal).

In the word entry the function can be illustrated by citation:

<entry>

<form>

<orth>базановець [bazanovets]</orth>

<gram>м. [m.]</gram>

</form>

<sense>

<def>

<lat>Lysimachia</lat>, вербейник [loosestrife].</def>

<cit>

<quote>Базановец (т) трава востволие именуемая. [*Basanovets* (explanation) herb called *vostvolije*.]</quote>

<bibl>

<title>Азбуковник Давида Замарая [*Azbukovnik* of David Zamaray]</title>

<date when='1626'>1626</date>

</bibl>

</cit>

<function>

<type>

<label>медицинская [medical]</label>

<cit>

<quote>Базановець, трава лѣкарская, нарицаемая востволие. [*Bazanovets*, medical herb called *vostvoliye*]</quote>

<bibl>

<title>Азбуковник Давида Замарая [*Azbukovnik* of David Zamaray]</title>

<date when='1632'>1632</date>

</bibl>

</cit>

</type>

</function>

</sense>

</entry>

Derivative forms will be attached to the word from which they were derived in the corresponding meaning. For example, the adjective *яблонный* 'apple (adj.)' will be given in the word entry *яблонь* 'apple tree':

<entry>

<form>

<orth>яблонь [yablɔnʹ]</orth>

<gram>ж. [f.]</gram>

</form>

<sense>

<def><lat>Malus sp. P. Mill. </lat>, яблоня [apple tree].</def>

<cit>

<quote>Ту абие томъ часѣ, добро и на потребу сыи древо масличное прозябло, и с нимъ и орѣхъ, и смокы, и яблонь. [Instantly at this moment, a good olive tree, usable for consumption sprouted there, and with it also a walnut tree, a fig tree, and an apple tree.]</quote>

<bibl>

<author>Иоанн экзарх Болгарский [John the Exarch]</author>

<title>Шестоднев [Shestodnev]</title>

<date notBefore='1451' notAfter='1475'>1451-1475</date>

<origDate notBefore='901' notAfter='920'>901-920</origDate>

</bibl>

</cit>

<cit>

<quote>И дрeвеса много овощнаа стоятъ бес числа, масличие, смокви, и рожцы, и яблони, и черешни, инородия. [And there are many fruit trees, standing innumerable, olive trees, fig trees, and carob trees, and apple trees, and cherry trees, others.]</quote>

<bibl>

<title>Хождение игумена Даниила [The Pilgrimage of Abbot Daniel]</title>

<date when='1495'>1495</date>

<origDate notBefore='1106' notAfter='1108'>1106-1108</origDate>

</bibl>

</cit>

<derivative>

<form>

<orth>яблонный [apple]</orth>

</form>

<cit>

<quote>Прививки яблонные. [Apple grafting.]</quote>

<bibl>

<title>Назиратель [Nazirateł]</title>

<num>333r</num>

<date notBefore='1576' notAfter='1600'>1576-1600</date>

</bibl>

</cit>

<cit>

<quote>Сады яблонные. [Apple gardens.]</quote>

<bibl>

<title>Роспись Китайскому государству... [Description of the Chinese country...]</title>

<date when='1619'>1619</date>

</bibl>

</cit>

</derivative>

</sense>

</entry>

Words denoting parts of the plants, such as bud, branch, wood, leave, grain, fruit, bark, root, flower, petal, sepal, bulb, oil, seed, sap, resin, cone, etc. will be represented in the separate word entries, for example:

<entry>

<form>

<orth>изюмъ [*iz'um*]/</orth>

<gram>м. [m.]/</gram>

</form>

<etym>от <lang> тур. [Turkish]/</lang> <foreign>üzüm</foreign>

</etym>

<sense>

<def>Плод растения <lat>Vitis vinifera L. </lat>, виноградина [Fruit of Vitis vinifera L., a grape].</def>

<cit>

<quote> Которые ж подаютъ овощи терпкие и мокрые лучши плодятся в долинахъ нежели на горахъ, для тое причины смокви, изюмы и иные овощи благоуханные болши по горамъ родятся. [Those vegetables which they serve, tart and juicy, bear fruit better in the valleys than in the mountains, for the same reason figs, grapes and other fragrant vegetables grow usually in the mountains.]</quote>

<bibl>

<title>Назиратель [*Naziratel*]/</title>

<num>119v</num>

<date notBefore='1576' notAfter='1600'>1576-1600</date>

</bibl>

</cit>

<derivative>

<form>

<orth>изюмный [*izumny*]/</orth>

</form>

<cit>

<quote>Изюмныя и винныя ягоды, и виноград на все стороны лопатами мечут. [Grapes and fig berries in all directions are thrown by shovels.]</quote>

<bibl>

<title>Сказание о роскошном житии и веселии [Story about luxurious life and joy]/</title>

<num>41</num>

<date notBefore='1600' notAfter='1700'> XVII в.</date>

</bibl>

</cit>

In fact, the dictionary should consist of two parts. The first part, “Phytonym — Plant”, or Russian — Latin, is organised according to the Russian alphabet, has Old Russian phytonyms as head entries, and explains which plants could be named by this or that Old Russian word, as it was shown in the given entries. The second part, “Plant — Phytonym”, or Latin — Russian, is organised according to the Latin alphabet, having modern nomenclature Latin plant names as word entry titles, and explaining which Old Russian names this or that plant could have in the 11th–17th cc. Some plant names, mostly loaned, could have quite a large amount of synonyms. For example, the database contains 10 words for peach (*Prunus persica* (L.) Stokes) — *бросква, броскви́на, броскви́ня, броски́ня, персика, персиконъ, прасква, родакина, шаптала, шептола*, and 10 words for sycamore fig (*Ficus sycomorus* L.) — *икамориумъ, каморнумъ, сокомория, сукамина, сукамня, сукомария, сюкамина, суюкомория, ягодина, ягодичина*. The abundance of variants reflects unstable situation in the lexical system caused by the appearance of the new objects and dialect variety.

The two-fold structure of the dictionary will allow making fast and comfortable cross-references between plants and their names.

Conclusion

The electronic format of the dictionary allows several people to work simultaneously, which is very important in a joint project. It also helps with representing new data in real time and easily preparing printed versions if needed. The Old Russian plant names dictionary is intended to be an open resource useful for both ethnobotanists and specialists in the Slavic historical lexicology. It will also become the basis for the research into the later periods of the Russian language.

The research is supported by the RFBR (the Russian Foundation for Basic Research), project 17-06-00376 “Russian Phytonyms in the Diachronic Aspect (11–17 cc.)”.

References

- Dictionary 11th–14th — *Slovar' drevnerusskogo jazyka (XI–XIV vv.)*. In 10 vol. Moscow: Russkij jazyk; 1988—. [Dictionary of the Old Russian language (11th–14th cc.)]
- Dictionary 11th–17th — *Slovar' russkogo jazyka XI–XVII vv.* Moscow: Nauka; 1975—. [Dictionary of the Russian language of the 11th–17th cc.)]
- Dictionary 16th–17th — *Slovar' obihodnogo russkogo jazyka Moskovskoj Rusi XVI–XVII vv.* St Petersburg: Nauka; 2004—. [Dictionary of Quotidian Russian of Muscovite Rus of the 16th–17th cc.]
- Kolosova V., Zaytseva K., Kovalenko K. (2018) PhytoLex — the Database of Russian Phytonyms: from Idea to Implementation in *SlaviCorp 2018. 24–26 September 2018. Charles University, Prague. Book of Abstracts*. P. 88–90. (https://slavicorp.ff.cuni.cz/wp-content/uploads/sites/144/2018/09/SlaviCorp2018_Book_of_Abstracts.pdf)
- 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 (<https://www.lexonomy.eu/docs/elex2017.pdf>)
- Prokhorov, G.M. (1997) Khozhdenie igumena Daniila [Abbot Daniel' Journey] (text edition, translation and comments by G.M. Prokhorov) in *Biblioteka literatury Drevney Rusi*. Vol. 4 (<http://lib.pushkinskijdom.ru/Default.aspx?tabid=4934>)

Roskov Y., Ower G., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T., DeWalt R.E., Decock W., Nieukerken E. van, Zarucchi J., Penev L., eds. (2019). *Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist*. Digital resource at www.catalogueoflife.org/annual-checklist/2019. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X. (<http://www.catalogueoflife.org>)

DICTIONARIES, CORPORA AND ARCHAIC WORDS: THE CHANGE OF CHINESE CHARACTERS WITH THE WOMAN RADICAL

Lan Li

The Chinese University of Hong Kong (Shenzhen)

Abstract

In line with the rapid change of the world, many new words are created to fill the gaps between language and reality, just as many die out. There are many studies on neologisms, and their inclusion in dictionaries, but very limited studies are about lexicographic practice with archaic words. How are these words treated? Are they still in dictionaries today? This study takes a particular group of Chinese words as an example, the words with the female character 女(*nü*). A lexicographic approach together with corpus analysis was employed to explore the change in this group of words. Nine dictionaries published at different times were reviewed, and an online mega-corpus of 287 million Chinese characters was searched to see the gap between a dictionary wordlist and words used in real data. It is interesting to find that about 1200 archaic characters with the radical *woman* listed in ancient Chinese dictionaries have vanished due to the social and cultural changes such as a shift in attitudes towards women. The reasons are discussed from social, cultural and linguistic perspectives. About 250 words survived in modern Chinese dictionaries are analysed quantitatively and semantically. The importance of the study is to shed light on the understanding and treatment of obsolete words in a language.

Key Words: Dictionary, language form, archaic words, gender

1. Background

Language is an indicator of social norms which speakers place on others and themselves. Gender is an obvious sign of social practice in the language. Gender systems are used in approximately one-quarter of the world's languages. Indeed, gender in language has received a good deal of scholarly attention. Hellinger and BuBmann (2002) summarised four linguistic categories for the representation of women and men in different languages: grammatical gender, lexical gender, referential gender and social gender. Grammatical gender is reflected either by articles, adjectives or by gender ending nouns. Lexical gender refers to words whose gender is not overtly seen but embedded in semantic features, such as *sister*, *mother* and *uncle*. It relates to the property of extra-linguistic femaleness and maleness. Referential gender indicates a reference to 'female', 'male' or 'gender-indefinite'. In this category, the use of male nouns can be 'generic masculine' referring to both genders. A female noun could also be a gender-indefinite reference in some languages. The category of social gender refers to the socially imposed dichotomy of masculine and feminine roles and character traits. Deviations from such sex-role assumptions will often require clear formal markings, as in English, *female surgeon* or *male nurse*. This practice indicates that many languages can be assumed to have a male bias, irrespective of whether the language has grammatical gender or not (Hellinger and BuBmann, 2002:9).

The gender orientation in the Chinese language is very different from Romanised languages. The Chinese language does not have a grammatical gender form; that is, there are no inflexions in nouns, verbs or adjectives when referring to a female. Gender visibility in the language is to be found in three areas: word-structure, word order and semantic representation of gender in words, idiom and proverbs (Chan, 1997). The word structure representing a woman will be discussed in this paper.

Tang (1988) surveyed the Chinese lexicon for words containing the radical woman 女(*nǚ*) and classified the words into four semantic categories: (1) words relating to marriage or giving birth, (2) kinship terms and terms regarding family relationships, (3) words referring to beauty, and (4) derogatory words or words with negative connotations (Tang, 1988, p.62). Later studies also followed this classification (Ettener, 2002; Chan, 2002) but the claims had little support from authentic data. Only part of the words with the woman radical was discussed in Tang (1988) and Chan (2002), lack off comprehensive and systematic analysis. This paper will take a lexicographic and corpus approach to discuss gender orientation in Chinese. The lexicographic approach will review the words containing the female radical recorded in dictionaries published at different times. It will start with orthographic and semantic analysis of the words, compare their inclusions in dictionaries and map them into different semantic groups. The dictionaries selected includes four desk-top Chinese dictionaries published after 2010, two online Chinese dictionaries and three desk-top Chinese-English dictionaries. They will be discussed in the next section.

The corpus approach has long been used in lexicographic studies. Quantitative information can be obtained from a large amount of real data, and the analysis of concordance lines can show a word's collocation, colligation, semantic prosody and semantic preference (Sinclair, 2004). Using corpus helps us illustrate how the words with the woman radical are used in contemporary society. Dictionaries and an online Chinese mega-corpus of 267 million characters were used for the following research purposes:

- 1) To examine how gender identity is constructed and reflected through language use;
- 2) To examine how such use has been changed over time.

2. Forms of gender in English

Gender in a language is not only grammatical but also political. It serves as a social identity that points to both cultural universality and variation, connecting the concepts of grammatical gender and political correctness, that is, whether women are treated equally or not in language. A comparison between English and Chinese will help to illustrate gender marks in the form of a language.

In modern English, grammatical gender is mainly confined to bound morphemes of nouns and roots of compounds. Some morphemes, including affixes and roots, can carry a presumed gender. For instance, suffixes *-or* and *-ess* reveal straightforward masculinity and femininity respectively, such as *prince* and *princess*. Other examples are *hostess*, *heiress* and *actress*. With the advent of the feminist movement, many marked female forms have become outdated and have been replaced by male forms referring to either sex, e.g. *author*, *professor*, *actor*, *hero*, and *heir*. Women in English are also presented in compounds with the word *woman* as a free root morpheme such as *congresswoman*, *spacewoman*, *laywoman* and *newswoman*; they mostly indicate job titles or professional titles. With women playing the same roles as men in many fields in modern society, gender-specific words of such have also changed. A gender-neutral lexical morpheme — *person*, has been used as a substitute for distinctive gender roots to include both genders. The examples are *layperson*, *chairperson* and *businessperson*.

3. Gender orientation in the Chinese language

The Chinese writing system primarily started from pictographic forms, but these forms gradually developed ideographic and logographic overtones through various combinations. Many Chinese characters today are compound graphs with a section head known as 部首 (*bushou*, section head). There are different English translations for this term. Most dictionaries call it ‘the radical’; others name it ‘the signfic’ (Ettner, 2002:32) or ‘semantic stem’(Fan, 1996). This paper takes the term ‘radical’, following the practice of most Chinese English dictionaries. Another point worthy of clarification is 字(*zi*, character) and 词(*ci*, word) in Chinese. A *zi* is a single character and can be called a monosyllabic word. Some corpus studies found this group take about 54% of the total words in a Chinese corpus (Huang et al., 2002). A *ci* normally has two or more characters; the most common ones are disyllabic or quadrisyllabic. The commonly held (but also often challenged) assumption is that the linguistic word should be the most basic lexical unit (Hartmann, 2003). A lexical unit in Chinese can contain one or more characters; therefore a single character is also a word. When the term ‘woman-word’ is used in this paper, it refers to a character with the radical 女 (*nü*, woman).

3.1 Lexicographic analysis of the woman words in Chinese

Chinese dictionaries are either character-based, called *zidian* (literally character dictionary), or word-based, referred to as *cidian* (word dictionary). The dichotomy dictates the definition of lexical entries: characters are lexical entries in a dictionary of characters and words are lexical entries in a dictionary of words. In this study, the author examined characters with the radical woman in nine Chinese dictionaries (see Table 1). Four desk-top dictionaries are published after 2010 and are regarded as representatives of modern Chinese. Two online dictionaries were selected because they have the largest number of characters with the radical *woman* 女 (*nü*); they have no space limit and can provide a comparatively complete etymological record of the words. Three Chinese-English dictionaries were also reviewed because their criteria for selecting entries are more on the practical side, providing users with the words used in everyday life. Table 1 shows the number of *woman* words in the nine Chinese dictionaries:

Table 1 Words with the radical *woman* included in modern Chinese dictionaries

Title	Publisher	Year of publication	No of character with女(woman)
现代汉语词典 Modern Chinese Dictionary	Commercial Press	2010	234
辞源 Chinese Etymology	Commercial Press	2010	274
新华大字典	Commercial Press	2012	181

Xinhua Chinese Dictionary			
汉语大字典	Sichuan Press	2010	984
汉典 9 (网络) Online Chinese Dictionary 9	http://www.zdic.net/	2004	1540
在线新华词典 Online Xinhua Dictionary	http://xh.5156edu.com/		477
现代汉语词典 (英译本) Modern Chinese Dictionary (English)	Foreign Language Teaching and Research Press	2010	202
汉英大词典 A Comprehensive Chinese-English Dictionary	Shanghai Translation Press	2010	169
当代汉英词典 (网络版) Contemporary Chinese-English Online Dictionary	The Chinese University of Hong Kong Press	1999	139

The table shows a big difference in number. *The Online Chinese Dictionary 9* (OCD9) (<http://www.zdic.net/>) gathered Chinese characters from several ancient and modern dictionaries. It has 1540 characters with the radical 女 (*nü, woman*). They are mostly from *康熙字典* (*Kangxi Zidian, The Imperial Character Dictionary of Kangxi*), an authoritative dictionary completed in 1716. However, more than 50% of woman-words in OCD9 are labelled as 生僻字 (*shenpi zi, rare word*) with no meaning or no pronunciation.

In contrast, most modern Chinese dictionaries include only around 200 characters with *woman* radicals, which means over 1200 woman characters have been abandoned. One reason for this dramatic drop is that rare words are no longer used today, and another reason is word variation. Since a lexical unit typically represents what language users perceive as a single minimal form-meaning pair, it allows some variations in forms. “In Chinese orthography, the variations go beyond graphic variations of the same glyph in different (historical, regional, or typographic) conventions” (Huang, Li & Su, 2016:541). For example, the word 娄 (*lou, family name*) has nine variants in *Kangxi Dictionary*. In modern dictionaries, it takes only two forms: the simplified form 娄 and

complex/traditional form 婁. With the time passing by, many characters with the radical *woman* have become obsolete. This is in line with the Chinese language change; only about 6,000 characters out of the 47035 in *Kangxi Dictionary* are used today.

In a Chinese dictionary, a character (an orthographic unit and an equivalent of a conventionalized sociological word) is given an entry according to the radical it contains (and hence its conceptual classification). The entry contains a rough definition of the semantic meaning of the radical and its phonetic element to form a particular word. *Xiandai Hanyu Cidian* (Modern Chinese Dictionary) (7th edition) and *Xinhua Zidian* (Xinhua Dictionary) (11th edition) list 201 radicals indicating semantic meanings of water, wood, person, woman, earth, plant, food, treasure, vehicle etc. The component contributes to semantic meaning to the character, but this is not a hard and fast rule. The phonetic element usually adds sound value indicating the pronunciation, to some degree, of the medial and possibly final segments of the character. For example, the concept *water* shaped into the form of 氵 (*san dian shui*), is a semantic component of 江 (*jiang*, river), 河 (*he*, river), 湖 (*hu*, lake), 海 (*hai*, sea) etc. The radical *wood* 木 (*mu*) forms 桌 (*zhuo*, table), 椅 (*yi*, chair), 板 (*ban*, board) 柜 (*gui*, cupboard) and so forth.

The etymology and applications of characters that use the semantic stem or the radical *woman* demonstrate the linguistic precedents that subsequently confirm gender inequality and the development of patriarchy (Fan, 1996). Etner (2002) believes that Chinese orthography provides interesting clues about the nature of ancient Chinese social structure, as well as the attitudes and values concerning the status of women at various historical periods. He noticed that particular words like 始 (*shi*, ancestor, beginning) and 姓 (*xing*, surname), contain the semantic stem 女 (*nü*, woman), which can be taken as evidence to suggest that during the earliest formative period of the Chinese written language, ancient China may have been a matrilineal society. This assumption could explain why a large number of words contain the meaning of woman. In the dictionary, *Ciyua*, (Chinese Etymology) about 1500 Chinese characters have the radical *woman*, a very productive semantic form. For example in the word 妈 (*ma*, mother): the radical 女 (woman) is a semantic head which can form words of concepts, actions or relations in connection with woman, and the other half 马 (*ma*, horse) is the phonetic component indicating how the word is pronounced; it usually does not contribute to the word meaning.

Of the nine dictionaries listed in Table 1, *Xiandai Hanyu Cidian* (Modern Chinese Dictionary) (7th edition) was selected for semantic analysis of the characters with the radical *woman*. The dictionary was compiled by top linguists of the Institute of Linguistics of Chinese Academy of Social Science in the 1980s and has been revised regularly. It is regarded as ‘a historic milestone in dictionary publishing in China (Sheng, 2002: i). It is the most authoritative Chinese dictionary in China and had an accumulative circulation of 40 million by 2002. The author analyzed all the words with the radical *woman* in this dictionary.

The semantic analysis indicates that characters with the radical *woman* not only refer to a person, naming and kinship terms, but also represent the image and social position of women. The later have had semantic shifts and significant changes in vocabulary choice. That is, words describing women have been changed. Table 2 shows the classification of the 234 woman words in *Modern Chinese Dictionary* (7th edition.)

Table 2 Classification of characters with the radical *woman* in Chinese

Semantic field	No	%	Example (<i>pinyin</i> , meaning)
Kinship	21	11	妈 (<i>ma</i> , mother), 婆 (<i>po</i> , mother in law), 姐 (<i>jie</i> , sister), 妹 (<i>mei</i> , younger sister), 奶 (<i>nai</i> , grandma), 姑 (<i>gu</i> , aunt on father's side), 姨 (<i>yi</i> , aunt on mother's side)
Naming (family name and women's give name)	32	16.9	姚 (<i>yao</i>), 姜 (<i>jiang</i>), 娄 (<i>lou</i>), 姬 (<i>ji</i>), 娜 (<i>na</i>), 婷 (<i>ting</i>), 娟 (<i>juan</i>), 姝 (<i>zhu</i>)
Stage of women's life	11	5.8	婴 (<i>ying</i> , baby), 妮 (<i>ni</i> , little girl), 姑娘 (<i>guniang</i> , young girl), 妇 (<i>fu</i> , woman) 媪 (<i>ou</i> , old woman)
Action	36	19	嫁 (<i>jia</i> , wed), 婚 (<i>hun</i> , marry), 娶 (<i>qu</i> , marry), 娩 (<i>mian</i> , give birth to), 奸 (<i>jian</i> , rape) · 嫉妒 (<i>jidu</i> , envy), 娱 (<i>yu</i> , entertain), 嬉 (<i>xi</i> , play with) 嫖 (<i>niao</i> , flirt with)
Appreciating	39	20.6	好 (<i>hao</i> , good), 娇 (<i>jiao</i> , effeminate), 妙 (<i>miao</i> , wonderful), 婉 (<i>wan</i> , graceful), 妩媚 (<i>wumei</i> , charming), 婷 (<i>ting</i> , graceful), 妥 (<i>tuo</i> , proper, sound)
Depreciating	13	6.9	奸 (<i>jian</i> , wicked), 嫌 (<i>xian</i> , dislike), 妖 (<i>yao</i> , demon) · 婪 (<i>lan</i> , greedy), 嫉妒 (<i>jidu</i> , jalous), 媸 (<i>qi</i> , ugly)
Godess	7	3.7	媿 (<i>wei</i>), 嫦娥 (<i>chang e</i> , a woman in the moon), 嫫祖 (<i>leizu</i> , god of silkworm), 媸 (<i>di</i> , god of washroom), 媸 (<i>ba</i> , ghost of draught)
Gender relation with the Empora	8	4.2	姬 (<i>ji</i>), 妃 (<i>fei</i>), 妲 (<i>dan</i>), 婉 (<i>wan</i>), 媛 (<i>yuan</i>) (Empora's women at different levels)
Abnormal gender relation	7	3.7	妾 (<i>qie</i> , concubine), 姘 (<i>pin</i> , mistress) 媸 (<i>chang</i> ,

			prostitute) 妓(<i>ji</i> , prostitute)
Job title	5	2.6	媒妁 (<i>meishuo</i> , matchmaker), 保姆 (<i>baomu</i> , babysitter), 婕妤(<i>jieyu</i> , manager of Empora's women)
Pronoun	2	1.1	她 (<i>ta</i> , she), 妳 (<i>ni</i> , female you)
Others	8	4.2	奴(<i>nu</i> , slave), 嫡(<i>di</i> , close paternal relation),
Total	234	100	

3.1.1 Female kinship terms

“In any language, lexical gender is an important parameter in the structure of kinship terminology” (Hellingger and Bubmann, 2002, p.8). China has traditionally drawn a sharp division between males and females and their roles in kinship relations. Chinese society was traditionally organised by a patrilineal, and patriarchal kinship system wherein the family name and family estate were handed down from generation to generation. The complex network of kinship terms is a manifestation of gender-differentiated vocabulary. It takes variables of gender, generation, and lineage into consideration, so that one differentiates one's mother from one's father or an older sister from a younger one, and so forth. Kinship terms take up 11% of the characters with the radical 女 in the dictionary.

Table 2 Woman kinship terms

Younger generation	Generation 1	Generation 2	Generation 3
妞 <i>niu</i> , daughter	姐 <i>jie</i> , elder sister	妈 <i>ma</i> , mother	奶 <i>nai</i> , patrilineal grandma
娃 <i>wa</i> , children	妹 <i>mei</i> , younger sister	娘 <i>niang</i> , mother	姥 <i>lao</i> , grandma on mother's side
	妻 <i>qi</i> , wife	婆 <i>po</i> , mother-in-law	
	媳妇 <i>xifu</i> , wife	姨 <i>yi</i> , aunt on mother's side	
	妯娌 <i>zhouli</i> , brother's wife	姑 <i>gu</i> , patrilineal aunt	
		婶 <i>shen</i> , uncle's wife	

Female kinship terms are frequently used in everyday language. With the gender marker, it is straightforward to see whether a relative is male or female. While many other words with the radical *woman* are reduced, this group will probably remain unchanged.

3.1.2 Action verbs

Action verbs with the radical *woman* have a limited number in Chinese, which may indicate limited roles or functions of women. The verbs relate only to marriage, entertainment, sexual acts and some

emotions. The semantic mark 女 (*nü*, female) shows the logic meaning components of the word; a woman must be involved in the process. It constitutes the initial part of verbs ‘to get pregnant’ 妊娠 (*renchen*) and ‘to give birth to a child’ 娩 (*mian*). Marriage verbs in Chinese vary in line with the agent of the verb. If a woman marries to a man, the word is 嫁 (*jià*, marry). The character consists of a radical 女 (*nü*, woman) and a phonetic 家 (*jià*, home), which means to give the woman a home. When a man marries a woman, the verb 娶 (*qǔ*, marry) is used. The two components are 取 (*qǔ*, take) and 女 (*nü*, woman). The meaning is that the man takes the woman, and the man and his family are supposed to shoulder most of the cost of the marriage and provide a house and a car for the new couple. The verb 结婚 (*jié hūn*, wed) is used when the subjects are man and woman. The word formation displays the social structure and attitude of marriage in China: the woman marries the man, and the man weds the woman.

Some verbs of entertainment also have a radical *woman*, indicating that women are an essential part of amusement. For example, 嬉 (*xī*, amuse), 耍 (*shuǎ*, play), 娱 (*yú*, entertain) 嬉 (*yáo*, game). Other two verbs with a man-woman-man structure, 嫖 (*piào*, to flirt with), or a woman-man-woman form, 嫖 (*nào*, to flirt with) imply sexual activities with men and women. 媾 (*gòu*, make love), 奸/姦/姦 (*jiān*, rape) are verbs for sexual acts and assaults. Most of these words are no longer used and will be discussed later in this paper.

Of the 36 verbs with the radical *woman*, 14 are pejorative. Mental verbs in this group mostly have negative connotations and are regarded by default as women’s typical problems, for example, to envy (嫉妒 *jídù*, 媚 *mào*), to hesitate (媿 *an*), to disturb (撩 *liào*), to dislike (嫌 *xian*), and to hinder (妨 *fang*). These words reflect the traditional concept of women’s behaviour in Chinese society, or rather, the remnants of Confucius idea :

唯女子与小人为难养也。

wei nuzi yu xiaoren wei nanyang ye

only woman and small person be difficult support also

Women and uneducated people are the most difficult to deal with.

Some of the negative words are no longer used; others have lost their gender indication and are used for both men and women — for example, the word 嫌 (*xian*; dislike) and 妨 (*fang* hinder). We randomly selected 100 concordance lines of the verb 嫌 (*xian*; dislike) from the mega-corpus *Chinese Web* and found that only 25 % of the sentences used women as the agents of the verb 嫌 (*xian*; dislike); the rest are men, institutions, people, even countries.

3.1.3 Words describing women

Words describing women seem to be polarised. Of the 234 words with the radical *woman* in *Modern Chinese Dictionary*, 39 positively describe women’s appearance, behaviour and attitude. Only 13 are negative words. This finding differs from other research which claimed that more than 20% of words containing the radical 女 (*nü*, woman) are derogative (Ettner, 2002; Tang, 1988). This is probably because I investigated all the word with the woman radical in the dictionary. The words describing the beauty or virtue of women highlight different aspects. The majority describes the good appearance of women such as 姘 (*hua*, beautiful), 姘 (*jiào*, pretty), 姘 (*yan*, stunning), 姘 (*hua*, exquisite), 媿 (*mei*,

angelic), 娉 (*fu*, gorgeous). Some refer to the figure as 妖 (*fou*, slim); others stress the manner 婥 (*chuo*, elegant), 媠 (*chuo*, clean and tidy), 媠 (*gui*, quiet and nice), 媠 (*yi*, friendly), 娜 (*na*, delicate). These words may help to create the message that women are ‘entirely constituted by the gaze of man’ (Williamson, 1985:80). In this sense, Chinese character formations and their attendant meanings can be said to contribute to gender bias (Xia & Miller, 2013). The negative words describe woman’s ugliness, such as 媠 (*chi*, ugly) and 媠 (*qi*, homely), or misbehaving as 妖 (*yao*, goblin), 媠 (*yin*, obscene), 媠 (*xie*, seducing) and 媠 (*jian*, wicked). On the whole, positive and neutral words take a bigger portion of the woman-words. There are also some words describing the wisdom of women such as 媠 (*ling*, smart), 媠 (*xian*, skilful), 媠 (*liao*, wise).

4. Corpus analysis of woman words in Chinese

The corpus analysis of *woman* words in Chinese aims to show how pervasive the words are used in modern society, and the image of women they generate. Previous studies claimed that about 20% of Chinese words containing the radical *woman* are derogative (Tang, 1988; Ettner, 2002), giving readers a negative impression of the woman words. To attest these arguments, we have explored a large amount of data online.

The data employed is a mega-corpus, the *Chinese Web* by the University of Leeds in the UK. It used the web claw technique and collected over 3.8 billion Chinese characters from 5 million URLs (Emerson, 2006). It is one of the most up-to-date Chinese corpora and can be searched freely on the website of Internet Corpora (<http://corpus.leeds.ac.uk/internet.html>). The online version of the *Chinese Web*, *Internet-zh*, has 281,660,631 tokens and 1,268,440 types or unique words. By comparing the frequencies of characters with the radical *woman*, we can get a picture of how these words are used in modern society. Table 4 lists the frequency of some words with the radical *woman*. The frequencies of the words have been normalised to the occurrence in one million words, written as ‘item per million’ (ipm).

Table 4. Top 50 characters with the radical *woman* in the *China Web* corpus

No	item	Freq (ipm)							
			7	婚 (<i>hun</i> , marriage)	98.5		12	姓 (<i>xing</i> , family name)	62.4
1	她 (<i>ta she</i>)	3646.3		姑 (<i>gu</i> , patrilineal aunt)	90.02		13	媠 (<i>yu</i> , fun)	57.8
2	好 (<i>hao</i> , good)	2657.9	8				14	娘 (<i>niang</i> , mother)	45.3
3	始 (<i>shi</i> , start)	650		姑娘 (<i>guniang</i> , girl)	90		15	娜 (<i>na</i> , name)	38.5
4	妈 (<i>ma</i> , mother)	287	9				16	媠 (<i>yi</i> , aunt)	37.09
5	媠 (<i>fu</i> , woman)	168.3	10	妹 (<i>mei</i> , younger sister)	81.66		17	媠 (<i>jiao</i> , beautiful)	31.2
6	媠 (<i>qi</i> , wife)	99.56	11	奶 (<i>nai</i> , patrilineal grandma)	68.25				

18	嫌 (<i>xian</i> , dislike)	28.1	35	耍 (<i>shua</i> , play)	14.8
19	委 (<i>wei</i> ,	27.7	36	奸 (<i>jian</i> , wicked)	14.6
20	娶 (<i>qu</i> , marry)	26.3		妆 (<i>zhuang</i> , makeup)	13.81
21	妙 (<i>miao</i> , wonderful)	24.8	37		
22	嫁 (<i>jia</i> , marry)	24.4	38	姜 (<i>jiang</i> , name)	13.3
23	妓 (<i>ji</i> , prostitute)	24.2	39	保姆 (<i>baomu</i> , nanny)	13.3
24	婴 (<i>ying</i> , baby)	23.5	40	姆 (<i>mu</i> , name)	13.2
25	姐 (<i>jie</i> , elder sister)	23.37	41	姥 (<i>lao</i> , grandma)	12.9
26	奴 (<i>nu</i> , slave)	22.5	42	姚 (<i>yao</i> , name)	12.7
27	娃 (<i>wa</i> , kid)	20.7	43	嫉妒 (<i>jidu</i> , envy)	11.3
28	婆 (<i>po</i> , mother-in-law)	20.5	44	娟 (<i>juan</i> , name)	10.1
29	媳 (<i>xi</i> , daughter-in-law)	20.1	45	嫂 (<i>sao</i> , sister-in-law)	9.1
30	妞 (<i>niu</i> , girl)	19.6	46	婪 (<i>lan</i> , greedy)	8.9
31	娇 (<i>jiao</i> , effeminate)	18.5	47	妥 (<i>tuo</i> , proper)	8.77
32	妨 (<i>fang</i> , hinder)	17.8	48	媚 (<i>mei</i> , charming)	8.5
33	嫩 (<i>nun</i> , tender)	16.3	49	婷 (<i>ting</i> , graceful)	8.2
34	妖 (<i>yao</i> , demon)	15.6	50	妨碍 (<i>fang ai</i> , hinder)	8.1

Word frequency lists have long been a part of the standard methodology for exploiting corpora. Sinclair stressed the prominence of frequency lists in corpus study. He noted, "anyone studying a text is likely to need to know how often each different word form occurs in it" (Sinclair, 1991, p.30). Comparing occurrences of the individual words in a text provides insightful information about the importance of words in the language. The higher the frequency, the more commonly the word is used. The table shows that the most frequently used word with the radical *woman* is a pronoun, 她(*ta*, she). Personal pronouns are grammatical words and always rank high in a word frequency list. The second is 好(*hao*, good), probably the most basic adjective in many languages. A dramatic fall in frequency can be seen after the first two characters. The third character 始 in line has 650 occurrences in 281 million words, only a fourth of the second character 好 with a frequency of 2658 ipm.

The statistic check of the 234 characters with the radical woman in *Modern Chinese Dictionary* revealed some unexpected results. 86 characters have a frequency of zero, which means they do not occur at all in the mega-corpus of 281 million words. This indicates that the number of gender-specific words in Chinese is reducing and a 'modern' dictionary should not include those words. The words with zero occurrence include the ones describing women such as 姘(*hua*, beautiful), 姣(*jiao*, charming), 妍(*yan*), 嬋(*hua*), 媵(*mei*), 婁(*fu*) 妖(*fou*, slim) 嫵(*chuo*, elegant), 媿(*chuo*, clean and tidy), 婉(*gui*, quiet and nice), 嫵(*yi*, friendly). Some gender-specific action verbs, for instance, 媿(*chuo*, arrange), 嫵(*liao*, cherish) 嫵(*liao*, disturb) are also found out of use. Some of these words have been replaced by characters without marked gender. For example, 淫(*yin*, obscene) has replaced 媿(*yin*, obscene), and 懶(*lan*, lazy) has taken over 媿(*lan*, lazy). **Their semantic heads were changed, but the phonetic components remained the same.**

5. Conclusion

Gender visibility can be easily seen in the Chinese language with the radical character *woman* 女(*nü*) forming a major component of many Chinese characters relating to women. The semantic analysis revealed that the meanings of these gender-marked words cover a limited range of semantic fields: kinship, naming, sexual relations, appreciating women, devaluing women. Gender disparity can also be conveyed in word order, compounds and different idioms. With drastic changes in society particularly regarding the role of women, many female-marked words, especially negative words and naming words have become redundant. The meanings of some words are now expressed in gender-indefinite characters. The corpus-based study has shown the decreased use of words with the radical *woman*. The kinship terms remain more or less the same, but the ones used for women's names and jobs are disappearing. Negative words with the woman radical are used to describe both genders.

The contribution of this study is the extensive survey of words with the radical *woman* and answered the question what changes have taken place in this particular group of words in modern Chinese. It enriches our understanding of gender profiles in the language, and may also benefit Chinese learners both at home and abroad. The author suggests Chinese lexicographers watch the language change and control the number of words with the radical woman when compiling a Chinese dictionary or a Chinese-English dictionary.

References

Baron, A., Rayson, P. & Archer, D. (2009). Word frequency and keyword statistics in historical corpus linguistics. *Anglistik*, 20 (1): 41-67.

- Chan, M. K. (1997). Gender differences in the Chinese language: A preliminary report". In: *Proceedings of the 9th North American Conference on Chinese Linguistics*, ed. by Hua, L. (pp.35-52). Los Angeles: GSIL Publications, University of California.
- Corbett, G. (1994). Gender and gender systems. In *The encyclopedia of language and linguistics*, ed. by R. Asher, (pp.1347-1353). Oxford: Pergamon Press.
- Curzan, A. (2003). *Gender shifts in the history of English*. Cambridge: Cambridge University Press.
- Emerson, T. & O'Neil, J. (2006). Experience building a large corpus for Chinese lexicon construction. In *Wacky! Working papers on the Web as Corpus*, ed. by Baroni, M & Bernardini, S. (pp. 41-62). Bologna: GEDIT.
- Ettner, C. (2002). In Chinese, men and women are equal – or – women and men are equal? In *Gender across languages: The linguistic representation of women and men*, ed. by Hellinger, M. & BuBmann, H. (pp. 29-56). Amsterdam: John Benjamins.
- Fan, C. (1996). Language, culture and Chinese gender. *International Journal of Politics, Culture and Society*, 10 (1), 94-114.
- Hartmann, R. (2003). *Lexicography: Reference works across time, space, and languages*. London: Taylor and Francis.
- Hellinger, M. & BuBmann, H. 2002. *Gender across languages: The linguistic representation of women and men*. Amsterdam: John Benjamins.
- Huang, C.R., Chen, C. R., & Shen, C. C. (2002). The nature of categorical ambiguity and its implications for language processing: a corpus-based study of Mandarin Chinese. In Mineharu, N. (ed.) *Sentence processing in East Asian languages* (pp. 109-116). Stanford: CSLI.
- Huang, C., Li, L. & Su, S. (2016). Lexicography in the contemporary period. In Chan, S.W. (ed.) *The Routledge encyclopaedia of the Chinese language* (pp. 540-566). London: Routledge.
- Lakoff, R. (1975). *Language and woman's place*. New York: Harper and Row.
- Nielsen, E. (1998). "Linguistic sexism in business writing textbooks". *Journal of Advanced Composition*, 8(1), 55-65.
- Prewitt-Freilino, J., Caswell, T. and Laakso, E. (2012). The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66, 268–281. DOI 10.1007/s11199-011-0083-5.
- Sheng, J. X. (2002). Preface. In *The contemporary Chinese Dictionary* [Chinese English edition] (pp. i-iii). Beijing: Foreign Language Teaching and Research Press.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language corpus and discourse*. London: Routledge.
- Tang, T. (1988). The phenomenon of stressing the importance of males and treating light the females in the Chinese lexicon. In *Monographs on modern linguistics studies on Chinese* ed. by Tang, T. (pp. 59–65). Taipei: Student Books.
- Williamson, J. (1985). *Decoding advertisements. Ideology and meaning in advertising*. London & New York: Marion Boyars.
- Xia, X., & Miller, R. E. (2013). Reconstructing gender ideologies of English loanwords in Chinese. *Language & Communication*, 33, 214–220.

Xiandai Hanyu Cidian (*Modern Chinese Dictionary*) (7th ed.) (2016). Beijing: Commercial Press.

Zhang, H. (2002). Reality and representation: Social control and gender relations in Mandarin Chinese proverbs. In *Gender across languages: The linguistic representation of women and men*, ed. by Hellinger, M. & BuBmann, H. (pp. 73-78). Amsterdam: John Benjamins.

AUGMENTING CROSS-LINGUAL TERMINOLOGIES WITH TREE-TO-SEQUENCE NEURAL MACHINE TRANSLATION

Long-Huei Chen, Kyo Kageura

The University of Tokyo

Abstract

Terminologies facilitate communication among domain experts. During knowledge exchange, term forms need to be consistent and context-independent to maintain the integrity of the underlying conceptual system. As such, term banks (collections of cross-lingual, cross-domain terminologies) are essential for correct term use across languages in the translation pipeline of humans, so automatic translation of terms remains a hot topic in research. Machine translation in general does not produce the exactness in forms required in term translation. To adopt neural translation methods for terminologies, we recognize that terms in different languages often exhibit common internal structures as they are essentially designations of shared concepts. To this end we design an architecture that encodes the terms word-by-word with a Tree-LSTM unit. The representation of compound terms (which account for 75-80% of all terminologies) is learned from dependency tree structures on top of the word unit embeddings of a particular language terminology. We test our model with cross-lingual term pairs from the Inter-Active Terminology for Europe (IATE), with English as the source language and French, Italian Spanish and Irish as target languages. The F1 and Exact Match scores in term translations are consistently higher by 1-3% in our model compared to the existing methods. F1 and Exact Match (EM) scores are applicable metrics in our study as terminologies have little room for paraphrasing. Our results are also significantly better than the Google Cloud Translation API trained on general corpora. The contribution of our work is twofold. First, in line with recent findings, we show that higher-quality data (in our case terminologies), in addition to the sheer quantity, are essential in the performances in language application tasks can be improved. Secondly, our work speaks to recent trends of including a more curated learning source aimed at specific target features in the end-to-end architectures.

Key Words: Multilingual Terminology, Machine Translation, Tree-structured Neural Network, Literal Translation

Augmenting Cross-lingual Terminologies with Tree-to-Sequence Neural Machine Translation

Introduction

Terminologies facilitate communication among domain experts. They are of importance during knowledge exchange to ensure that a compatible set of terms are applied to describe the underlying conceptual systems (Sager, 1990). Application-wise, term banks (collections of cross-lingual, cross-domain terminologies) are essential in the translation pipeline of human translators, as they connect correct term use across languages.

The Inter-Active Terminology for Europe (IATE) (Johnson & Macphail, 2000) is the official term bank sanctioned by the European Union (EU). It is the go-to source of terminologies for translators working with the official European Union languages. Since its creation in 1999, IATE serves the needs to standardize

term use throughout all EU institutions, and today contains approximately 1.4 million multilingual entries submitted by all of the EU’s translation services.

A persistent problem with many term banks, however, is that the difference in abundance among languages. In IATE we especially see a gap between languages recognized by EU earlier and those added later. Since terminologies are of little use to translators when they lack cross-lingual data, translating terminologies between languages remains a hot topic in terminological research (Navigli, Velardi, & Gangemi, 2003; Itagaki, Aikawa, & He, 2007; Nikoulina & Dymetman, 2015; Farajian, Bertoldi, Negri, Turchi, & Federico, 2018). The task is defined as providing an accurate translation of specialized terms in some language based on term presence in another higher-resource language.

Method

Translating Terminologies

Term translation methods often borrow from existing lexicon induction research and attempt to bring terms into another language by mapping words across languages. But since 75-80% of terms are multi-worded (Kageura, 2012), word ordering for terms in another language often requires extra care (Tsvetkov & Dyer, 2016; Sharoff, 2018). On the other hand, conventional phrase-based machine translation approaches are less applicable to terms, since they are built from phrasal units of a larger scale than terms.

Terminology translation requires precision in forms, as there is little room for paraphrasing for terms. As a result, existing translation models trained on textual corpora are often not up to par when translating terminologies. Here we test with the public Google Cloud Translation API1, and the F1 and the Exact Match (EM) scores between the translation and the gold standard show huge room of improvement. We observe that the system produces translations that are close in meaning but considered incorrect use in the terminology.

Tree-to-Sequence Term Translation

As terms are essentially designation of concepts (Sager, 1990), terminologies often exhibit internal relationship structures not typically expected of general vocabularies. This structure is supposedly shared across languages, as the head-modifier structure of multi-worded terms reflects the knowledge discovery and evolution process that motivates term creation. Thus the term translation task can be construed as a mapping of terminological structure between languages, as they all sprang from the same underlying knowledge representation.

We propose to learn a hidden representation of the terminological structure by encoding the terms with a tree-LSTM built on top of a term’s dependency parse tree. The output from the tree-encoder is then combined with the hidden state from a conventional sequential encoder, and passed onto the decoder towards our goal of augmenting a terminology with terms from another language.

Child-Sum Tree-LSTM Encoder

We construct a tree-based encoder with long short-term memory (LSTM) cell units, where each node in the dependency parse tree is represented with an LSTM unit (See Figure 1). Given a tree, let $C(j)$ denote the set of children of node j . The Child-Sum tree-LSTM transitions are modified from LSTM cells and described in the following equations (Tai, Socher, & Manning, 2015):

$$\begin{aligned}
\tilde{h}_j &= \sum_{k \in \mathcal{C}(j)} h_k \\
i_j &= \sigma(W^{(i)}x_j + U^{(i)}\tilde{h}_j + b^{(i)}) \\
f_{jk} &= \sigma(W^{(f)}x_j + U^{(f)}h_k + b^{(f)}) \\
o_j &= \sigma(W^{(o)}x_j + U^{(o)}\tilde{h}_j + b^{(o)}) \\
u_j &= \tanh(W^{(u)}x_j + U^{(u)}\tilde{h}_j + b^{(u)}) \\
c_j &= i_j \odot u_j + \sum_{k \in \mathcal{C}(j)} f_{jk} \odot c_k \\
h_j &= o_j \odot \tanh(c_j)
\end{aligned}$$

Intuitively, we consider the parameters in this tree model as representations of correlation between the components of the tree-LSTM unit, the input x_j , and the hidden states h_k of the unit’s children. The attentional mechanism is also applied on top of each of the word unit in the terms, but not on the tree-internal LSTM units, which differs from the approaches in (Eriguchi, Hashimoto, & Tsuruoka, 2016). Also, we incorporate input-feeding (Luong, Pham, & Manning, 2015) by feeding the previous unit of word prediction into the current decoder hidden unit (Figure 2).

Decoder and Initial Settings

We have two encoded states, one from a tree-based encoder as described and the other from a conventional sequential encoder as in (Luong et al., 2015). To generate the initial state s_0 in the decoder, we make a new tree-LSTM unit which has as children the final hidden states from both the sequential and the tree encoder (Zoph & Knight, 2016):

$$s_0 = g_{tree}(h_n, h_{root}^{(tree)})$$

Results

We train and compare among three models:

an attentional encoder-decoder Recurrent Neural Network (RNN) model with the tree + sequential two-way encoder from Section

an attentional encoder-decoder RNN model with a conventional sequential encoder (Luong et al., 2015)

a multi-head attention Transformer model (Vaswani et al., 2017)

Data Source

We select five representative language pairs of different data sizes (Table 1) from the Inter-Active Terminology for Europe (IATE) to build the train set. The test set is created as a holdout with 10% of the data size in each language; we train on the train set and report the test set evaluation results.

Model and Training Details

The vanilla-sequential model has two layers in the encoder to make the model complexity comparable among all three models. The Transformer contains 12 heads with a 2048 hidden feed-forward size. The

input for each word is built from the 300-dimensional pretrained multi-lingual fastText embeddings (Bojanowski, Grave, Joulin, & Mikolov, 2016), and during the training process we allow the model to back-propagate into the embeddings to enable embedding learning. We set the learning rate at 0.001 with the Adam optimizer, and train for 20000 steps with the learning rate decayed by half after half the training steps.

Term Translation by Language

For each language pairs, an individual model is trained to compare the performance when data sizes differ. The F1 and Exact Match (EM) scores per language are shown in Table 3.

Term Translation by Domain

Terms in IATE are also divided into knowledge domains. For English-to-French term translation, we experiment with models trained separately for each domain, whose details is given in Table 2. The results are in Figure 3.

Discussion

We find that the results are consistently higher across languages and domains when we allow for tree + sequential two-way encoders to learn from term structures (Table 1 and Figure 3). The complexity with the multi-head attention Transformers means that they are hard to be useful considering the nature of our data. Combined these results suggest that our tree-LSTM architecture can learn common terminological structures (which are motivated by knowledge structure across languages) from the term data, and thus contribute to the performance of our proposed model.

We also find that the results do not correlate strongly with data sizes across both languages and domains; this means that the internal consistency of dependency structures is as critical as the data size. On the other hand, we are unable to find that limiting model training to single domain data boosts performance, which suggests that the internal structure of terms should not be restricted to single domains to increase data efficiency.

Related Work

Bilingual Term Induction

Mikolov et al. (Mikolov, Le, & Sutskever, 2013) first noted that word embedding spaces exhibit common structures across languages, when they are exploited to linearly map word embeddings from a source to a target language. Several studies aimed at improving these cross-lingual word embeddings (Faruqui et al., 2014; Xing, Wang, Liu, & Lin, 2015; Ammar et al., 2016). The same techniques are applied to terminologies or specialized vocabularies (Sharoff, 2018).

Neural Machine Translation

Kalchbrenner and Blunsom (Kalchbrenner & Blunsom, 2013) were the first to propose an end-to-end Neural Machine Translation model. The model is solidified with an encoder-decoder architecture with GRU RNNs (Cho et al., 2014) and LSTMs (Sutskever, Vinyals, & Le, 2014). Bahdanau et al. (Bahdanau, Cho, & Bengio, 2014) and Luong et al. (Luong et al., 2015) proposed and refined the attention mechanism NMT so that it can dynamically focus on local windows. Multi-head self-attention mechanisms are proposed (Vaswani et al., 2017) and raised state of the art in neural machine translation.

Tree-LSTM and Tree-to-sequence Translation

Tree-LSTM modifies the original LSTM cells that enable a rich network topology in which each LSTM unit can take in information from multiple child LSTM units. The architecture is used when the sentence representations are crucial (Tai et al., 2015). Eriguchi et al. (Eriguchi et al., 2016) applied tree-LSTM to neural machine translation (NMT) by running the source through the Head-driven Phrase Structure Grammar (HPSG) to get a binary tree with multiple phrase unit forming a sentence structure (Sag, Wasow, Bender, & Sag, 1999).

Conclusion

The contribution of our work is twofold. First, we show that by using high-quality data (in our case terminologies), performances in language application tasks can be improved. In the context of terminology processing and extraction tasks, our observation is in line with recent findings that data quality is at least as important as the sheer quantity of the data (Morin, Daille, Takeuchi, & Kageura, 2010; Zhang, Gao, & Ciravegna, 2018; Spasić, Corcoran, Gagarin, & Buerki, 2018). Secondly, our work speaks to recent trends of including a more curated learning source in the end-to-end architecture (Chen et al., 2018; Birch, Finch, Luong, Neubig, & Oda, 2018). We show that progress can be secured when the learning process is aimed at specific target features.

Unlike multilingual translation models trained on general corpora, we demonstrate that our terminology-oriented model design can contribute concretely to the augmentation of terms across languages. We are currently refining our approach to further take into account the information defined by the paradigmatic structure of terminologies.

References

- Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016). Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Birch, A., Finch, A., Luong, M.-T., Neubig, G., & Oda, Y. (2018). Findings of the second workshop on neural machine translation and generation. arXiv preprint arXiv:1806.02940.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
- Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., . . . others (2018). The best of both worlds: Combining recent advances in neural machine translation. arXiv preprint arXiv:1804.09849.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
- Eriguchi, A., Hashimoto, K., & Tsuruoka, Y. (2016). Tree-to-sequence attentional neural machine translation. arXiv preprint arXiv:1603.06075.
- Farajian, M. A., Bertoldi, N., Negri, M., Turchi, M., & Federico, M. (2018). Evaluation of terminology translation in instance-based neural mt adaptation.

- Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166.
- Itagaki, M., Aikawa, T., & He, X. (2007). Automatic validation of terminology translation consistency with statistical method. Proceedings of MT summit XI, 269–274.
- Johnson, I., & Macphail, A. (2000). Iate–inter-agency terminology exchange: Development of a single central terminology database for the institutions and agencies of the european union. In Proceedings of the workshop on terminology resources and computation, Irec 2000 conference. athènes, grèce.
- Kageura, K. (2012). The quantitative analysis of the dynamics and structure of terminologies (Vol. 15). John Benjamins Publishing.
- Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1700–1709).
- Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.
- Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.
- Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2010). Brains, not brawn: The use of smart comparable corpora in bilingual terminology mining. ACM Transactions on Speech and Language Processing, 7(1), 1.
- Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. IEEE Intelligent systems, 18(1), 22–31.
- Nikoulina, V., & Dymetman, M. (2015, February 5). Terminology verification systems and methods for machine translation services for domain-specific texts. Google Patents. (US Patent App. 13/955,315)
- Sag, I. A., Wasow, T., Bender, E. M., & Sag, I. A. (1999). Syntactic theory: A formal introduction (Vol. 92). Center for the Study of Language and Information Stanford, CA. Sager, J. C. (1990). Practical course in terminology processing. John Benjamins Publishing.
- Sharoff, S. (2018). Language adaptation experiments via cross-lingual embeddings for related languages. In Irec 2018 proceedings.
- Spasić, I., Corcoran, P., Gagarin, A., & Buerki, A. (2018). Head to head: Semantic similarity of multi-word terms. IEEE Access, 6, 20545–20557.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104–3112).
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.
- Tsvetkov, Y., & Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. Journal of Artificial Intelligence Research, 55, 63–93.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 1006–1011).

Zhang, Z., Gao, J., & Ciravegna, F. (2018). Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 57.

Zoph, B., & Knight, K. (2016). Multi-source neural translation. arXiv preprint arXiv:1601.00710.

Footnotes

¹We chose the <https://cloud.google.com/translate/docs/Google Cloud Translation API> as it's a popular API with support for all EU languages.

Tables

Table 1

Total number of bilingual terms pairs with respect to source and target language pairs selected.

Source	Target	Term Pairs
English	French	585569
English	Italian	380217
English	Spanish	371993
English	Irish	58203

Table 2

Number of English-to-French terms pairs in the five selected domains.

Index	Domain	Term Pairs
28	Social	106087
68	Industry	88964
48	Transport	79244
4	Politics	59222
32	Education	51279

Table 3

F1 and Exact Match (EM) scores among baseline and proposed models for the target language pairs.

Target Language	French		Italian		Spanish		Irish	
Scores (%)	F1	EM	F1	EM	F1	EM	F1	EM
(Luong et al., 2015)	40.17	16.23	40.72	15.96	45.08	18.17	37.39	15.75
(Vaswani et al., 2017)	29.81	8.26	29.78	9.03	12.79	1.55	18.65	3.30
Ours	41.85	16.96	41.46	16.74	45.58	18.75	38.36	16.20

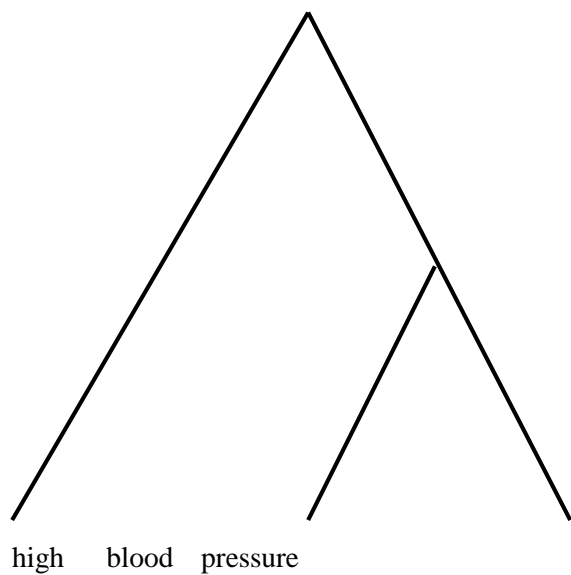


Figure 1. Dependency parse tree of a sample term high blood pressure, which shows how the tree-LSTM encodes the term.

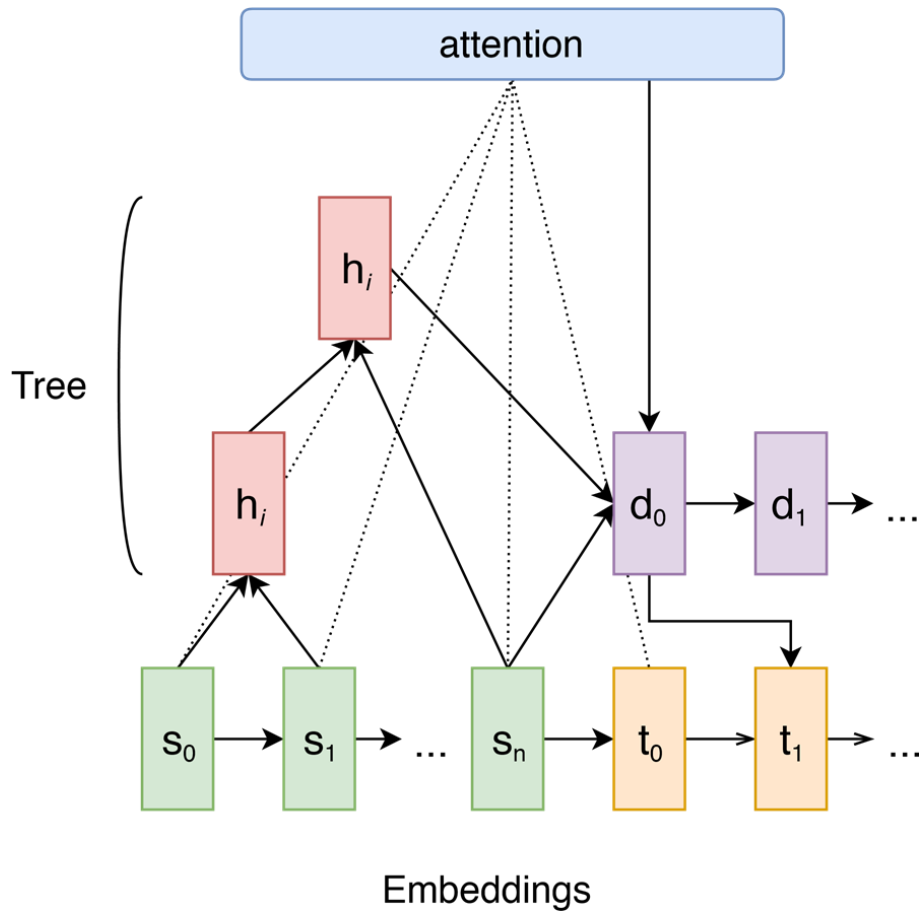


Figure 2. The proposed tree-to-sequence attentional encoder-decoder RNN for term translation. A tree-LSTM hidden state is combined with the sequential encoder output, and passed along to the decoder to incorporate structural information.

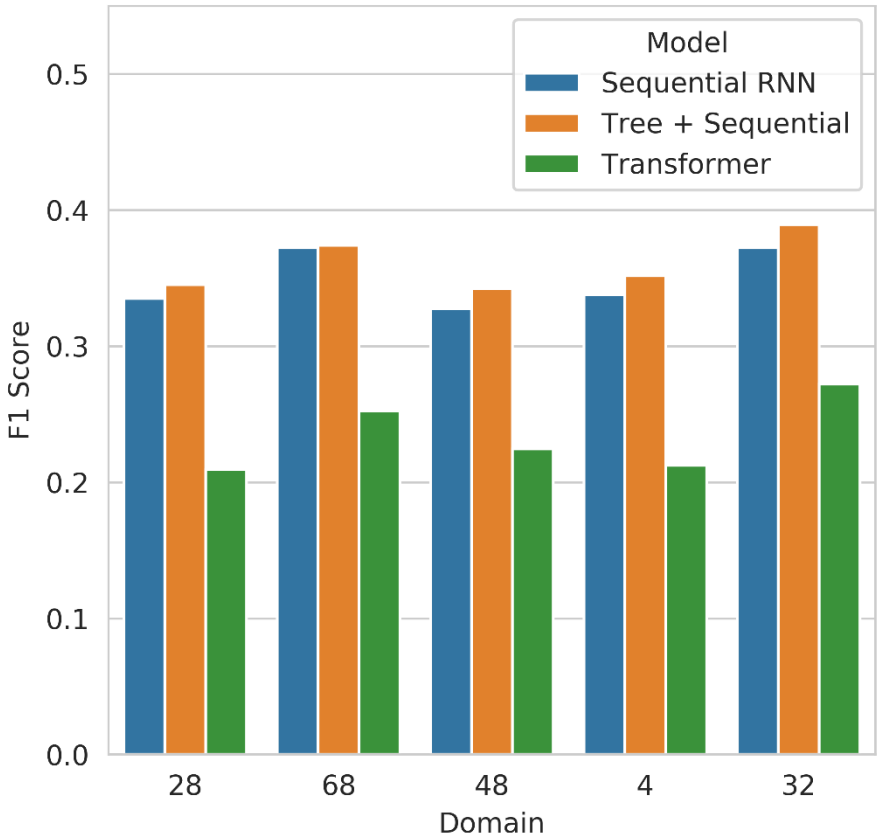


Figure 3. Results for English-to-French translation when models are trained individually for by-domain term pairs. We see that F1 scores for the tree + sequential two-way model are consistently higher across all domains despite the difference in data sizes.

THE ROLE OF CORPORA AND E-LEXICOGRAPHY IN THE DIDACTICS OF FRENCH PHRASEOLOGY

ALBANO Mariangela

University Sorbonne Nouvelle, Paris 3 & University Dokuz Eylül of Izmir

Laboratory HTL (Histoire des Théories Linguistiques - UMR 7597, Université Sorbonne
Nouvelle – Paris 3 et Université Paris Diderot – Paris 7)

Abstract

This paper analyses the role of corpus data and online dictionaries to teach French multi-word units to adult students of French as Foreign Language (Sinclair 1991 & 1998; Chi 2003 & 2011; Jantunen 2016; Prinsloo et al. 2017; Paquot in press).

In L2 didactics it is important to stress the difficulties a learner can run into in both comprehension and translation of various multi-word units (Lewis 2000; Wray 2002; Granger & Meunier 2008; González Rey 2010). The frozenness and the figurative dimensions characterising them can be easily interpreted by native speakers (Wood 2006; Pawley & Syder 2000). However, L2 learners usually do not have direct access to phraseological units meaning (Cacciari & Tabossi 1988; Gibbs 1986) and memorizing them requires a preliminary work of morphosyntactic patterns recognition (Waara 2004; Cavalla & Labre 2009; Boers et al. 2010; Benigni et al. 2015).

Faced with the plurality of approaches on didactics of phraseology, we took into consideration the hypothesis of proposing some didactical strategies in French language for adult students of French as Foreign Language. These didactic units are based on the integration of the cognitive approach (Lakoff & Johnson 1980, 1999; Langacker 1987; Fillmore et al. 1988; Croft 2001; Tomasello 2003), the motivation theory approach (Dobrovolskij 1995, 1997, 2004; Dobrovolskij & Piirainen 2005), and the analogy theory approach (Gentner 1983; Hofstadter 1995; Gentner et al. 2001).

This kind of approach allows us to show how the conceptualization and the metacognitive strategies are stimulated in the comprehension and the translation of fixed expressions through corpora and e-dictionaries.

Key Words: online corpora, e-lexicography, multi-word units, French phraseodidactics, cognitive linguistics.

1. Introduction

A learner is a social subject who organizes language into meaningful networks based on his mental representations, his pre-knowledge and the context in which he learns. This means that learning a language results from the integration of a social, cognitive and cultural dimension. From this perspective the didactics of phraseology should take into account the possibility of encouraging learning by reflecting on the cultural-motivational dynamics of frozen expressions, by an 'exercise' of metacognitive thinking and by cognitive strategies to interpret a fixed expression.

Accessing to cultural chunks means to take into account the main tool reflecting a culture, that is the "dictionary" (Ridel 2008: 2). Indeed, since the beginning of the lexicographic research in French context (Quemada 1968; Dubois 1970; Rey 2007) and, according to Pruvost (2006) and Galisson R. (1988), which support a lexicultural approach, a dictionary is an intermediary that allows access to a situated *doxa*.

In recent years, next to this tool, large corpora have changed the possibilities for accessing to a culture (Gross M. 1975; Sinclair 1991 & 1998; Hausmann 2004). In particular, lexical and grammatical forms, and the different occurrences of a headword give access to a culture through the linguistic uses.

Moreover, according to several theoreticians (Sinclair 1991 & 1998; Chi 2003 & 2011; Jantunen 2016; Prinsloo et al. 2017; Paquot in press) the adoption of these means in a foreign language classroom could help to make effective the educational action of vocabulary activities. In particular, phraseodidactics could benefit from this kind of approach because cultural symbols in language or, according to Pamies (2007) “*culturèmes*”, are reflected in fixed chunks (Piirainen 1998 & 2010; Dobrovolskij 1998; Dobrovolskij & Piirainen 2005a & 2005b).

From this perspective, this article aims to show how corpus data and online dictionaries could help to teach French multi-word units to adult students of French as Foreign Language. More specifically, we describe the methodology applied in the construction of didactics units of French frozen expressions (i.e. idioms, collocations, locutions, proverbs, phraseme, idiomatic enunciates). This methodology was applied two times: 1) in 2018, in the third year of B.A., in a French Semantics for translators class with foreign speakers and Erasmus students (University Sorbonne Nouvelle, Paris 3)⁴⁷; 2) in 2019, in the second year of B.A., in French as foreign language class with Turkish mother-tongue students (University Dokuz Eylül).

Our theoretical positioning is embedded in cognitive (Lakoff & Johnson 1980 & 1999; Kövecses & Szabó 1996; Pütz et al. 2001; Boers et al. 2010) and metacognitive (O'Malley & Chamot 1990; Cardona, 2001) approaches.

We also try to harmonize the methodologies derived from lexical approach (Lewis 1993 & 2001) and construction grammars (Langacker 1987; Fillmore, Kay & O'Connor 1988; Tomasello & Brooks 1999; Croft 2001; Waara in Achard & Niemeier 2004; Wee 2007; Delorme Benites in Perrin & Kleinberger 2017; Benigni et al. 2015) with work on motivation (Dobrovolskij 1995, 1997, 2004; Dobrovolskij & Piirainen 2005a/b) and analogy (Gentner 1983; Cacciari & Tabossi 1988; Cacciari & Levorato 1989; Holyoak & Thagard 1989; Cacciari & Glucksberg 1991; Hofstadter 1995; Gentner, Holyoak & Kokinov 2001).

In Didactics of French as Foreign Language, for example, thanks to Bally's legacy (1951 [1909]: I, II), numerous attempts at didactisation of fixed expressions were produced either on French-speaking ground (Binon & Verlinde 2003; Cavalla & Crozier 2005; Cavalla et al. 2005; Tutin 2007; Granger & Meunier 2008; Cavalla et al. 2009; Legallois & Tutin 2013; Cavalla 2014) and on non-French speaking ground (Lüger 1997 & 2004; Ettinger & Stölting 1992; Ettinger 1992; González Rey 2002, 2007, 2010).

Faced with this plurality of works, we took into consideration the hypothesis of constructing some didactic units of French frozen expressions integrating some exercises proposed in the domain of French language teaching with the previous theoretical framework.

This kind of approach not only allows us to better understand the interpretation process of a fixed expression, but also makes it possible to identify how could we teach it.

For this article, we have chosen to show how we construct and we submit to students two didactic units concerning French fixed expressions.

Even if the number of examples will not be exhaustive to represent the whole of the results, we try here to propose a preliminary study of the hypotheses on the didactisation of fixed expressions in order to give, on the one hand, tools to improve their teaching.

⁴⁷ In University Sorbonne Nouvelle, the class of French Semantics for translation is obligatory for the third year of B.A.. The majority of students were Spanish mother-tongue studying Spanish and English; a little group speaks French as L2 and studied Spanish and English; there was a little number of French mother-tongue that studied English and Spanish or English and Russian; others were Italian studying French and English; Chinese students studying French and English; Bulgarian students studying Russian and French.

Our article has two parts: firstly, we show our methodology introducing the steps: 1) explicit teaching of morphology, lexicology and semantics; 2) explicit teaching of lexicography (in particular, dictionaries and corpora); 3) explicit teaching of languages for specific purposes; 4) preparation of an individual work on fixed expressions ; 5) construction of an individual video-phraseologism. Secondly, we examine the results of our didactical units in the different groups.

2. Methodology

Our study is based on different steps: 1) explicit teaching of morphology, lexicology and semantics; 2) explicit teaching of lexicography (in particular, dictionaries and corpora); 3) explicit teaching of languages for specific purposes; 4) preparation of an individual work on fixed expressions ; 5) construction of an individual video-phraseologism.

2.1 The construction of metalinguistic competence through explicit teaching of morphology, lexicology and semantics

For building a metalinguistic competence in L2, students have to keep a cognitive distance from the language, to assume it as «object of thought» (Pinto & El Euch, 2015: 7). Indeed, according to Benveniste (1974 : 228-229), «the metalinguistic faculty refers to the possibility that we have to rise above the language, to abstract from it, to contemplate it, while using it in our reasonings and observations»⁴⁸.

Teachers, in their role of facilitators, can induce students to develop an analytical language knowledge through explicit teaching of linguistics. From this perspective, some researchers (Norris & Ortega, 2000; Ellis, 2005; Ortega, 2009; Cunningham, 2015) show that explicit teaching, where metalinguistic rules are explained to learners, could develop an enhancement of the learning strategies.

In our study, we borrow the concept of “shared metacognition” from Sagnier (2010)⁴⁹, by specifying that, in the case of the treatment of frozen expressions, students activate, in the first place, metalinguistic knowledge and, more specifically, meta-phonological, meta-syntactic, meta-semantic, meta-discursive and meta-pragmatic skills.

In the second place, they mobilize the “meta-social cognition”; in other words, the social meta-representations that could result from the use of a fixed expression in the mother tongue that allows the learner to reflect on the differences with the foreign language and take distances from the social group in which he/she is involved.

That being stated, for our study we worked on explicit teaching of morphology, lexicology and semantics.

First, we focused our attention on the segmentation level in the linguistics domain mentioning the morphological level, the lexicological level and the semantic level.

After that, we insisted on the definition of “morphology” and the differences between lexical morphemes and grammatical morphemes giving some examples and some exercises to the students. In particular, we consider very useful for the morphology didactics the simplified classification made by Monneret (2009 [1999]: 113)⁵⁰.

Moreover, we think useful for students to introduce some concepts related to the lexical morphology such as “affixation”, “derivation” and “composition” (Riegel, Pellat & Rioul 2014 [1994]) and other morphological

⁴⁸ Orig. fr. « la faculté métalinguistique renvoie à la possibilité que nous avons de nous élever au-dessus de la langue, de nous en abstraire, de la contempler, tout en l'utilisant dans nos raisonnements et nos observations ».

⁴⁹ This concept affects the following activities: activation of metalinguistic knowledge and meta-social cognition, planning, selective attention, memory, problem identification, and elaboration (Sagnier, 2010).

⁵⁰ Indeed, the author specifies some examples of lexical morphemes (fr. lexèmes) showing the difference between free (such as, nouns and adjectives) and bound lexemes (such as, lexemes radicals), and some examples of grammatical morphemes (fr. grammèmes) where he suggests three typologies: 1) free grammatical morpheme (i.e. preposition, adverb, conjunction, pronoun, determinant); 2) bound grammatical morpheme (i.e. verbal radical, non-verbal radical, derivational morphemes, flexional morphemes); 3) blended grammatical morpheme (such as the French preposition “de + le” becoming “du”).

processes such as synapse and different forms of abbreviation (Benveniste 1966: t.2, 175; Mortureux 2013 [2004]).

Concerning the domain of lexicology, we try to initiate students to the concepts of “word” and “sentence” explicating the difficulty to establish a concrete difference (Saussure 1967 [1916]; Bally 1951 [1909]; Meillet 1952; Martinet 1966; Chiss, Filliolet & Maingueneau 2016).

In particular, we showed agglutinative and analogical processes (Saussure 1967 [1916]; Bally 1951 [1909]) working on polylexical units and fixed expressions.

We added a short description on neology and loan words processes according to Lemaire & Campenhoudt (2008).

We proposed to students some exercises to identify and recognize different kinds of words in their mother-tongue.

After that, we defined the concepts of connotation and denotation, of “phraseology” and we introduced the phraseological categories following different approaches (Cowie 1981; Gross G. 1996; Rey *in* Martins-Baltar 1997; Svensson 2004; Burger et al. 2007; González Rey 2008). In particular, we explicated the differences among idioms, locutions, collocations, phrasemes, proverbs, gallicismes and idiomatic enunciates.

Moreover, we added the linguistic phenomena characterizing lexicalization such as grammatical block (Heinz 1993; Hudson 1998), syntactical block (Fraser 1970; Fontenelle 1994; Nunberg et al. 1994; Gross G. 1996; Mejri 1998; Wray 2002), marked syntax (Gross M. 1984; Anscombe 1984; Gross G. 1993; Mejri 1998; Schapira 1999), refusal of the para-synonymic substitution (Martin 1976; Hudson 1998); unique context (Misri 1987b; Schapira 1999; Svensson 2004), polylexicality (Gross G. 1996), non-compositionality (Nunberg et al. 1994; Martin 1997; Moon 1998), semantic opacity (Newmark 1988; Gibbs 1994; Gross G. 1996; Moon 1998; Schapira 1999) and conventionality (Nunberg et al. 1994; Moon 1998).

According to different researchers (Bally, 1951 [1909] : I,II; Ettinger, 2014b), we announce very briefly some information about etymology connecting this subject with the lexical semantics. Indeed, we explored semantic relations such as synonymy, polysemy, antonymy, homonymy, hyperonymy and hyponymy proposing some exercises of identification of semantic relations.

After that, we introduced the concept of lexical fields as Cavalla suggested for building a didactic unit (Cavalla & Crozier, 2005; Cavalla et al., 2005; Cavalla et al., 2009) and metaphorical motivation (Lakoff, 1987; Sweetser, 1990 ; Bybee, 2003 ; Brinton & Traugott, 2005 ; Boers & Lindstromberg, 2008). For doing this, we took into account the cognitive semantics (Lakoff & Johnson 1980, 1999; Langacker 1987; Fillmore et al. 1988; Croft 2001; Tomasello 2003; Dobrovolskij 1995, 1997, 2004; Dobrovolskij & Piirainen 2005).

In particular, according to Lakoff & Johnson (1980 & 1999), we introduced the notion of conceptual metaphor; this is to say the relation between a conceptual starting domain, defined "source domain", to an arrival domain, defined as "target domain". However, this projection produces a set of epistemic correspondences or mappings concerning a conceptual system (Lakoff & Johnson, 1980: 3).

For giving access to different kinds of daily-life metaphors, we offered to the students a list of the most important conceptual metaphors: orientational metaphors (i.e. happy is up; sad is down; conscious is up; unconscious is down), ontological metaphors (i.e. inflation is an entity; the mind is a machine; container metaphor), and structural metaphors (i.e. argument is war) (Ibid.).

We also take into consideration the primary metaphors that, according to Grady et al. (1996), concern the connection between our sensory-motor experience to the domain of our subjective judgments (i.e. intimacy is proximity, difficulties are weights, categories are containers, similarity is closeness, change is movement, seeing is touching); the basic metaphors such as death is a final destination, people are plants (Lakoff & Turner, 1989); the generic metaphors following generic structures such as basic ontological categories (i.e. entities,

states, events, actions or situations) or aspects of being (attributes, behaviors) or the figure of events (cyclicity, instantaneity or extension, singularity and repetition, destructive or creative entities) (Lakoff & Johnson, 1980 and 1999); and complex metaphors (i.e. union of different primary metaphors).

Having stressed that idioms fit conceptual metaphors and are part of the daily-life language, we introduced the concept of semantic motivation. In particular, motivation can be presented under the intra-linguistic prism as in the case of the English word «cupboard», which is more motivated by its written form than by its phonological structure (Boers & Lindstromberg, 2008: 18). It can also be intrinsically linked to physical, social and cultural experience and, in this case, the motivation will be extra-linguistic.

Motivation has been discussed in cognitive linguistics either synchronically (Lakoff & Johnson, 1980; Kövecses, 1986; Boers, 1999; Jäkel, 2003) or diachronically (Sweetser 1990; Bybee 2003; Brinton & Traugott 2005). In particular, it is worth noting that in the case of fixed expressions, motivation is based on the embodied rooting relating to different domains such as emotions, economics, architecture, medicine or teaching (Lakoff & Johnson, 1980; Kövecses, 1986; Boers, 1999; Jäkel, 2003; Caballero, 2003a; Salager-Meyer, 1990; Low, 2003)⁵¹.

Applying motivation studies on didactics of phraseology, show the benefits that learners can gain by examining the motivation of fixed expressions. These benefits concern first of all the speed of access in the figurative sense thanks to this form of double mental decoding (form, signified); secondly, learning the lexicon using a vocabulary organized semantically and no longer with a list to memorize (Kövecses & Szabó, 1996; Kövecses in Pütz et al., 2001; Boers, 2000b); thirdly, the perception of the plausibility of motivations that makes the learning of a foreign language more concrete (Boers & Lindstromberg, 2008).

We also introduced the diachronic study of motivation based on the approach of Boers and Lindstromberg (2008) because learners not only can appreciate the nature of words and sentences, but they can also make etymological comparisons with their L1 (Ibid.: 24)⁵².

After having suggested these usage-based language theories, we introduced the concept of construction grammars (Fillmore, Kay & O'Connor, 1988; Croft, 2001; Croft & Cruse, 2004; Goldberg, 2006). In particular, we try to differentiate different concepts such as decoding idioms (i.e. kick the bucket), encoding idioms (i.e. answer the door), grammatical idioms (i.e. spill the beans), extra-grammatical idioms (i.e. first off), idioms with a pragmatic point (i.e. once upon a time), idioms without a pragmatic point (i.e. by and large), substantive idioms (i.e. kith and kin, all of sudden, pull someone's leg) and formal idioms (i.e. the Xer, the Xer, cousin three times removed).

Moreover, we proposed a schema where we highlighted the identification of idioms anomalies (such as, the strangeness of grammatical form or the idiom meaning not inferable from the words), the recognition of the basic construction through the research of his abstract form, the specification of the meanings and the pragmatics concerning the idiom.

⁵¹ From this perspective, Boers (2000b) obtained results on the treatment and learning of verbal collocations in English by French-speaking learners. He insists on the importance of teaching conceptual metaphors like *VISIBLE IS OUT* (find out, discover, unmask, turn out, turn off, cut, expel, empty) and *VISIBLE IS UP* (to look up, to raise the eyes, to improve, to seek, to show up, to stand out, to point out, to bring out, to humiliate) to facilitate the understanding of the collocations (Ibid.). Similarly, Kövecses and Szabó (1996) and Kövecses (2001) also took into account the conceptual metaphors, *MORE IS UP* and *HAPPY IS UP* to teach verbal collocations in English to Hungarian learners.

⁵² Benefits related to the study of diachronic motivation have been highlighted by Ilson (1983), Christiansen (1995), Boers, Demecheleer and Eyckmans (2004a and 2004b), Boers (2004). More specifically, Ilson (1983: 81) argues that this work can help disambiguate difficult idioms, associate the derived form with its source form; to memorize the units with mental images; increase 'affective' motivation and interest in the 'other' language by focusing on the culture and history of a foreign language.

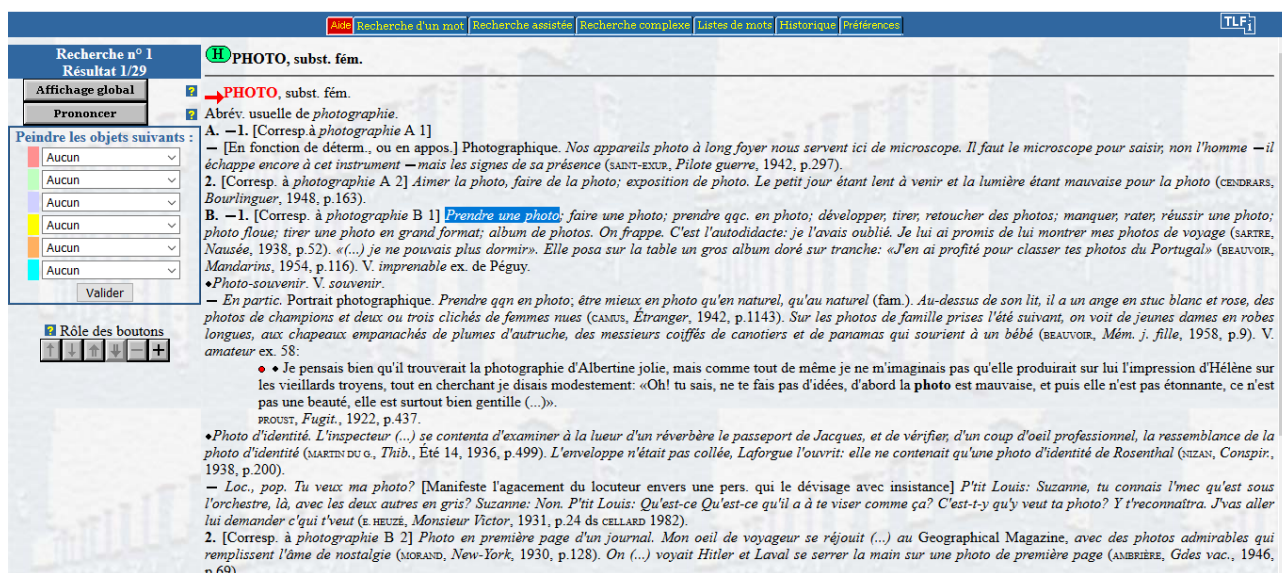
2.2 The didactics of lexicography through the attention on language for specific purposes

A special attention was paid to the presentation of the lexicography. In particular, we introduced the differences between monolingual and bilingual dictionaries and we paid attention on some concepts related to the terminology in the lexicographic domain.

Specifically, we analysed the concepts of “definition”, “lemma”, “entry”, “example” situated within a dictionary. We explicated the position of idioms and collocations in a dictionary entry giving some examples.

For doing this, we used the French online dictionary “Trésor de la Langue Française informatisée” (<http://atilf.atilf.fr/>) that we consider useful for different reasons: 1) being free and online, students could consult it everywhere; 2) being analytical and detailed, it gives us the possibility to examine all the parts of a lemma; 3) being rich of examples, it is a good tool for showing the fixed units in little contexts.

If we consider, for example, the French collocation “prendre une photo”, engl. “take a picture”, we could see above that a student could have the possibility to understand the meaning of this collocation through a lot of examples.



The screenshot shows the search results for the word "PHOTO, subst. fém." in the "Trésor de la Langue Française informatisée" dictionary. The interface includes a search bar at the top with the word "PHOTO" entered. Below the search bar, there are navigation options like "Recherche d'un mot", "Recherche assistée", "Recherche complexe", "Listes de mots", "Historique", and "Préférences". The main content area displays the word "PHOTO, subst. fém." followed by its definition: "Abrév. usuelle de photographie." and "A. -1. [Corresp. à photographie A 1] - [En fonction de déterm., ou en appos.] Photographique. Nos appareils photo à long foyer nous servent ici de microscope. Il faut le microscope pour saisir, non l'homme - il échappe encore à cet instrument - mais les signes de sa présence (SAINT-EXUP., Pilote guerre, 1942, p.297). 2. [Corresp. à photographie A 2] Aimer la photo, faire de la photo; exposition de photo. Le petit jour étant lent à venir et la lumière étant mauvaise pour la photo (CENDRARS, Bourlinguer, 1948, p.163). B. -1. [Corresp. à photographie B 1] Prendre une photo; faire une photo; prendre qqc. en photo; développer; tirer; retoucher des photos; manquer; rater; réussir une photo; photo floue; tirer une photo en grand format; album de photos. On frappe. C'est l'autodidacte: je l'avais oublié. Je lui ai promis de lui montrer mes photos de voyage (SARTRE, Nausée, 1938, p.52). «(...) je ne pouvais plus dormir». Elle posa sur la table un gros album doré sur tranche: «J'en ai profité pour classer tes photos du Portugal» (BEAUVOIR, Mandarins, 1954, p.116). V. imprenable ex. de Péguy. •Photo-souvenir. V. souvenir. - En partie. Portrait photographique. Prendre qqn en photo; être mieux en photo qu'en naturel, qu'au naturel (fam.). Au-dessus de son lit, il a un ange en stuc blanc et rose, des photos de champions et deux ou trois clichés de femmes nues (CAMUS, Étranger, 1942, p.1143). Sur les photos de famille prises l'été suivant, on voit de jeunes dames en robes longues, aux chapeaux empanachés de plumes d'autruche, des messieurs coiffés de canotiers et de panamas qui sourient à un bébé (BEAUVOIR, Mém. j. fille, 1958, p.9). V. amateur ex. 58: • Je pensais bien qu'il trouverait la photographie d'Albertine jolie, mais comme tout de même je ne m'imaginai pas qu'elle produirait sur lui l'impression d'Hélène sur les vieillards troyens, tout en cherchant je disais modestement: «Oh! tu sais, ne te fais pas d'idées, d'abord la photo est mauvaise, et puis elle n'est pas étonnante, ce n'est pas une beauté, elle est surtout bien gentille (...).» PROUST, Fugit., 1922, p.437. •Photo d'identité. L'inspecteur (...) se contenta d'examiner à la lueur d'un réverbère le passeport de Jacques, et de vérifier, d'un coup d'oeil professionnel, la ressemblance de la photo d'identité (MARTIN DU G., Thib., Été 14, 1936, p.499). L'enveloppe n'était pas collée, Laforgue l'ouvrit: elle ne contenait qu'une photo d'identité de Rosenthal (SUZAN, Conspir., 1938, p.200). - Loc., pop. Tu veux ma photo? [Manifeste l'agacement du locuteur envers une pers. qui le dévisage avec insistance] P'tit Louis: Suzanne, tu connais l'mec qu'est sous l'orchestre, là, avec les deux autres en gris? Suzanne: Non. P'tit Louis: Qu'est-ce qu'est-ce qu'il a à te viser comme ça? C'est-t-y qu'y veut ta photo? Y t'reconnaitra. J'vas aller lui demander c'qui t'veut (S. HEILZ, Monsieur Victor, 1931, p.24 ds CELLARD 1982). 2. [Corresp. à photographie B 2] Photo en première page d'un journal. Mon oeil de voyageur se réjouit (...) au Geographical Magazine, avec des photos admirables qui remplissent l'âme de nostalgie (STORAND, New-York, 1930, p.128). On (...) voyait Hitler et Laval se serrer la main sur une photo de première page (AMBIÈRE, Gdes vac., 1946, p.69).

Fig. 1

Some lessons were dedicated to the concordances, e-dictionaries and corpora and how to use them. More specifically, we used the list of texts created by Greaves (<https://www.lexutor.ca/conc/fr/>), the online dictionaries “Trésor de la Langue Française informatisée” (<http://atilf.atilf.fr/>), “Larousse” (<http://www.larousse.fr/dictionnaires/francais>) and, concerning the synonyms in French language, “Collins” (<http://dictionnaire.reverso.net/francais-synonymes/Collins>).

We found useful to introduce to the students the concept of corpora using the free online multilingual corpora created by the University of Leipzig (http://corpora.uni-leipzig.de/de?corpusId=fra_mixed_2012).

The Leipzig’s University corpora shows different advantages: 1) being free, students could access to it everywhere and they can work on it at home; 2) it is multilingual and students could search some expressions in their own languages; 3) there are a lot of texts called “Beispiele”, engl. “Examples”, and it is possible to extend results typing each time “+100”; 4) the section “Kookurrenzen” give to the students the possibility to deepen, through the number of occurrences, the meaning of a word; 5) the section called “Graph” show a lexical schema where the searched word is connected with other terms as we could see below:

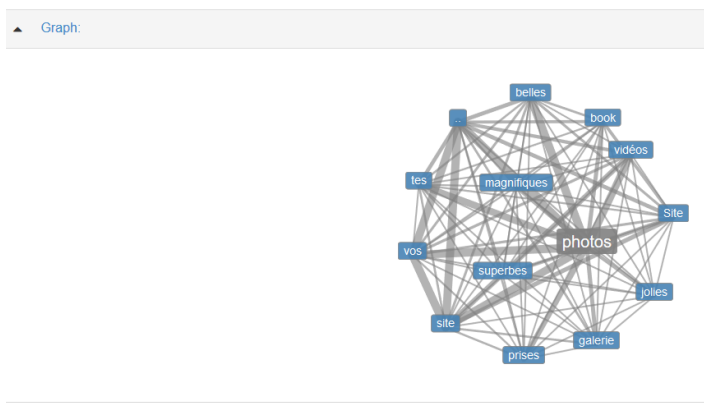


Fig. 2

Thus, a student that would like to know the meaning of the French collocation “prendre une photo”, engl. “to take a picture” could be helped by this schema which offers some lexical combinations.

To help students in their research of fixed words, we introduced the concept of “language for specific purposes” and we gave some examples in the domain of informatics, technology, medicine, law. For doing that, we consulted with the students some tools as we could see below:

- ATILF (Analyse et Traitement Informatique de la Langue Française), FRANTEXT, <https://www.frantext.fr/>
- Centre National de Ressources Textuelles et Lexicales, <http://www.cnrtl.fr>
- Club d’orthographe de Grenoble, Mots nouveaux des dictionnaires, Corpus DiCo (Dictionnaires comparés), <https://orthogrenoble.net/mots-nouveaux-dictionnaires>
- Koehn, Philipp, EuroParl (European Parliament Proceedings Parallel Corpus), <http://www.statmt.org/europarl/>
- Ortolang, Outils et Ressources pour un Traitement Optimisé de la LANGue, <https://www.ortolang.fr/> et <http://cnrtl.fr/>, 2015

2.3 How could learners analyze fixed expressions ? The construction of the phraseological entry-form

In order to improve the learning ability of phraseological expressions, we have given each student a text (i.e. newspapers, blogs pages, scientific articles, literary texts) characterized 1) by the presence of a specific language for the first group; 2) by the presence of cultural stereotypes for the second group.

Concerning the Sorbonne University group, we have proposed to students texts in medicine, psychology, economics, computer science, law because their French language level was B2/C1.

Then, students searched for five expressions per article and we asked them to fill in a form, prepared in advance, by us for each chosen expression.

Concerning the Dokuz Eylül University students showing a French language level equivalent to B1/B2, we have proposed some texts where they could extract one expression concerning cultural stereotypes and we asked them to fill the same form.

We called the proposed form, “the phraseological entry-form” because it gives to the students the possibility to detect all the particularities, the exceptions, the uses, the translation in their mother-tongue and in their L2 or L3 of a phraseological expression.

The phraseological entry-form could represent a personal syllabus that could help the students in different ways: 1) to push students to reflect on the «language» system, preparing them for an increasingly conscious, adequate and autonomous use; 2) to highlight any difficulties encountered in the translation and interpretation of a phraseologism and optimize time; 3) to integrate it with other exercises; 4) to develop a series of personal cards that the student can consult not only during the preparation of exams or tests, but also when he has finished his studies and in the domain of a translation job or a teaching unit.

The phraseological entry-form represents a tool that can be thought of both for the teacher, who can use it to explain a phraseme in the classroom, and for the student who can use it as an in-depth tool for the progression of his terminological skills.

For building the phraseological entry-form we took into consideration different approaches in phraseodidactics: the German school of phraseodidactics, and the studies within the FRAME project in Italy.

First, it is Kühn (1987) who introduces us to a pragmatic description of the phraseology that will be addressed in his teaching of German as Foreign Language. The author describes three fundamental steps for the didactics of frozen sequences that we can call 1) recognition and identification of frozen sequences; 2) understand the meaning; 3) use the frozen units.

The first is to make learners aware of the recognition of frozen units in an authentic text orally or in writing. Learners are urged to activate decoding processes by examining the structural properties of frozen expressions as well as for comparative frozen expressions (such as + adjective / adverb / noun / adverb), verbal complexes consisting of a functional verb and a nominal part. or a supplement (i.e. "Abstand von etwas / jdm nehmen", "to get away with someone / something") or for collocations (such as Zwillingsformeln in German "(auf) Biegen und Brechen", "at any price"). In this first phase, teachers promote learners to detect syntactic, morpho-syntactic and semantic abnormalities and to explore their use on the basis of communicative and situational contexts.

The second step involves deciphering the sense of frozen units within a communicative situation using a dictionary, personal reasoning or encouragement from the teacher. The usefulness of this phase also depends on a series of exercises proposed by the teacher to use phraseology in different contexts (Kühn, 1992: 182).

In this respect, Kühn (1994: 425) states that learners must learn to «to decode the context by non-phraseological substitutes and to resort to a typical situation, a specialist text and a pragmatic use related to the addressee».

Finally, the third step relates to the use of frozen units within contexts that are similar to those examined. It is a phase marked by an active role of the learner as suggested by our author:

Allen Schwierigkeiten zum Trotz sollten die Schüler im DaF [Deutsch als Fremdsprache] - Unterricht auch Hilfestellungen erwarten können Phraseologien regelgerecht zu gebrauchen. Gerade hier spielt die strikte Kontextualisierung aller Übungen eine entscheidende Rolle, denn der Schüler kann einen Phraseologismus nur dann regelgerecht verwenden, wenn er mit der Verwendungssituation vertraut ist. Der aktive Gebrauch von Phraseologismen sollte sich daher unbedingt auf die für den Lernenden nachvollziehbare Situationen und Kontexte beziehen (Kühn, 1994 : 425)⁵³.

Although the innovation brought by this method must be acknowledged, it has been improved by several theoreticians who have questioned the context (Ettinger, 1989); on the difficulty of the third stage for learners (Ettinger, 1998: 207), on the typologies of images specific to the fixed expressions to be taught (Heinz, 1993 and 1994), on which phraseological expressions to examine during a course, on the types of text, content or

⁵³ (I translate): "In spite of all the difficulties, students should also be able to wait for help in DaF (German as a Foreign Language) lessons to use phraseology correctly. Here, the strict contextualization of all the exercises plays a decisive role, because the student can use phraseology only if he is familiar with the use situation. The active use of phraseology should therefore necessarily refer to situations and contexts that the learner understands.

stylistic marks to be addressed (Hessky, 1992: 167) or the argumentative approaches of phraseology (Lüger, 1992, 1993, 1996, 1996a, 1997).

In 1998, Ettinger and Bárdosi, for example, constituted a "pragmatisation dossier" for German-speaking learners of French as Foreign Language where we can remark eleven exercises: 1) identification of the French phraseological unit by adding the quotation or the context and its source; 2) definition of the frozen expression identified in French; 3) illustration of the key concepts (using the studies of Bárdosi, Ettinger and Stölting, 1998); 4) adding the German phraseological unit or paraphrase in German; 5) indicating the typology of the fixed expression on the basis of the lists drawn up by Heinz (1993 and 1994); 6) recognizing grammatical restrictions concerning people, time, mode, active or passive voice, negation, interrogation, etc. ; 7) identifying the level of language or stylistic marks (i.e. literary, colloquial, popular, vulgar, slang, aged, old); 8) indicating the classificatory restrictions of the subject and the complements (i.e. name of thing, abstract noun, name of person, masculine, feminine, age); 9) describing the conditions of use of the frozen unit (who? To whom? When? Where? With what intention?); 10) playing the frozen expression through gestures or mimicry; 11) writing an etymological note on the identified expression.

Ettinger then added to this summary a description of the teacher's work during a university semester (Ettinger in Lorenz-Bourjot & Lüger, 2001). The author illustrates the main objectives which will be:

- 1) Einblicke zu geben in wesentliche Eigenschaften eines Phraseologismus : Polylexikalität, relative Fixiertheit, semantische Idiomatizität, Restriktionen grammatikalischer und klassematischer Art, Gebrauchsbedigungen (= *pragmatisation*);
- 2) Erwerb von ca. 500 Redewendungen, deren Kenntnis mit Hilfe zahlreicher zumeist formaler Übungen des Lehrbuches gefestigt wird ;
- 3) Mit Hilfe eines Arbeitsblattes sollen Möglichkeiten aufgezeigt werden, wie man selbständig weiterlernen könnte (*apprentissage autonome*) und wie man ganz konkret z.B. LE MONDE auf CD-ROM zum Erstellen einer eignen kleinen Sammlung auswerten könnte (*Ibid.* : 88-89)⁵⁴.

Later, Ettinger (2008: 108) proposed to follow five main steps to lead the learner to question the impossibility of paraphrasing a frozen unit and to allow him/her to understand this unit on the basis of the contexts of use.

The learner should therefore 1) select the frozen expression within a context; 2) illustrate the possible uses; 3) report syntactic, morphological, morpho-syntactic and classic restrictions; 4) explain the linguistic register and the pragmatic functions of the frozen sequences; 5) look for a key concept for each fixed expression using an onomasiological scheme.

Thanks to this, the learner will be able to develop two types of skills: on the one hand, the communicative one and, on the other hand, the phraseological one. As Solano Rodríguez points out (2004: 411),

El hecho de conocer las diferentes UF de una lengua, y saber interpretarlas e integrarlas en un discurso propio, oral o escrito, adecuado según el cotexto, el contexto, la relación con el interlocutor, las normas sociales y nuestros propósitos de interacción⁵⁵.

Different is the approach of his colleague Lüger, who in his first studies sketched the general lines to follow for a language phraseology (Lüger, 1997). In fact, he delineates the fields of intervention of this new discipline which are the following: 1) the semantic field where the learner is stimulated to understand the changes of meaning due to the frozen expression; 2) the contextual domain where the learner is encouraged to identify

⁵⁴ (I translate): "1) To give an overview of the essential characteristics of a phraseology: polylexicality, relative fixity, semantic idiomaticity, restrictions of a grammatical and classematic nature, conditions of use (= pragmatisation); 2) Acquisition of about 500 sentences whose knowledge is consolidated using many mainly formal exercises from the manual; 3) Using a worksheet, it is possible to show the possibilities of continuing to learn independently (autonomous learning) and to be very concrete in the creation of a small collection [of expressions frozen in help] eg [from] THE WORLD on CD-ROM "

⁵⁵ (I translate): "to know the different UF [fixed units] of a language, and to know how to interpret and integrate them in a correct speech, oral or written, appropriate according to the text, the context, the relation with interlocutor, social norms and our goals of interaction. "

and select frozen expressions in context; 3) the syntactic domain by which the learner can understand the notion of "restriction"; 4) the domain of paraphrasing, where the learner is urged to build the capacity to distinguish frozen structures from non-fixed structures by evaluating global meaning, compositionality and false friends; 5) the pragmatic domain that is needed to develop production skills (Ibid. : 85-89).

In our study, we referred to a second approach called "The Frame Project". In particular, the reflection on construction grammars in Italy allowed a group of theoreticians to form an international project entitled "FRAME" (FRaseologia Multilingue Elettronica, Fr. Multilingual Electronic Phrasology)⁵⁶, which currently plays a pioneering role in the development of a didactic of the frozen lexicon⁵⁷.

This project starts with the Schafroth's PhraseoFrame (2013, 2014a, 2014b) for Italian Foreign Language and is complemented by a pragmatism for other languages with the studies of Benigni et al. (2015), Imperiale and Schafroth (2016), Imperiale (2016). This is a multilingual database, not yet completed, where teachers and learners of a foreign language can find lists and descriptions of phraseology of the following languages: Italian, French, Chinese, English, German, Japanese, Russian and Spanish⁵⁸.

The approach used to present the phraseological elements is that of Goldberg's Grammar Construction (1995) and Croft's Grammar (2001), which provides a holistic description with several points of view: phonological, morpho-syntactic, semantic-pragmatic, and discursive. However, unlike Croft's approach, Benigni et al. (2015: 285) adopt a point of view that allows to «describe all the construction-specific and language-specific properties in a way that allows learners to understand and use these phrases»⁵⁹.

The database consists of two parts: the first, called "entry field" proposes a traditional definition of lemma that can be consulted quickly; the second, called "description field" offers a detailed description of the linguistic levels examined (morphological, syntactic, semantic-pragmatic and discursive).

In turn, the two parts present other sub-domains or "virtual spaces" and, in order not to burden the description of two parts, we will make the main part of the analysis on the elements that characterize them to the using two tables.

⁵⁶ The project website is: <http://www.fraseologia.it/gfm/>

⁵⁷ The universities involved in the financing of this project are the University of Milan, the University of Roma Tre and the Heinrich Heine University of Düsseldorf. We also want to highlight the presence of a project that has developed alongside FRAME and uses its data to build an even more advanced didactic tool. This is "Fraseodidattica multimediale" planned by Giacomina, Brunetti and Ruggieri in 2015 at the Technische Universität Dresden. This project, still in the initial phase, proposes the creation of a series of video-phraseologism. The authors have already created the first video-phraseologism for German-speaking learners of Italian LE and this sequence is entitled "Sono alla frutta! », Fr. "I'm at dessert!". Giacomina (2012 and forthcoming), who has worked extensively in the field of phraseology, emphasized the benefits of video-based phraseology because it affirms the importance of highlighting all the pragmatic aspects that contribute to the enunciation of a frozen expression.

⁵⁸ It should be noted that at present the metalanguage of the explanatory parts of the database is Italian because, according to the terms of Benigni et al. (2015: 282) "the 'ideal' learner we address is an Italian student, whose level ranges from A1 to C2", fr. The 'ideal' learner to whom we are addressing is an Italian student whose levels vary from A1 to C2 '. However, the authors state that there is the possibility of translating all entries into the language under consideration, that is to say that a phraseology in French will be explained in French and so on.

⁵⁹ In addition, the authors chose to implement the database using an onomasiological approach. As they say very clearly (Ibid., 2015: 285): «We work on the different semantic fields, looking for the most representative phraseological units of a given field in each language; analogies and differences will result from a comparative query. Therefore, the aim is not to find equivalents in a "target" language, but to describe thoroughly the phraseological units [...] The initial concepts on which we are working at the moment are 'greetings' and 'pleasures of the table' and at present we have about 50 units described. To complete the first step of the project, we plan to take about 10-15 phraseological units for each semantic field».

Field of entry		
	Domains ou Virtual Spaces	Description
1.	Phraseologism text field	This is an area where the learner or teacher can find texts and auditory materials for each language consulted.
2.	Transcription field	This domain concerns languages with alphabetic systems different from Latin.
3.	Transliteration field	This domain concerns languages with alphabetic systems different from Latin.
4.	Variant field	This is a section where the learner or teacher can examine the variants of the unit when they exist. The variants can be represented by minimal variations at lexical, morphological and syntactic level. In addition, the variants are organized according to diastratic, diatopic and diaphasic axes which present the following options: colloquial, vulgar, regional, formal, archaic and jargon.
5.	Literal translation	The learner or the teacher can consult the Italian literal translation of the frozen unit in a foreign language.
6.	Equivalent field	It is a virtual space dedicated to equivalents.
7.	Type of phraseologism field	In this section, the phraseological element is classified as an idiomatic expression or schematic idiom / construction phrasing (a proverb, a collocation, a formula).
8.	Formality field	It is a space divided into two levels: the first presents the degrees of formality (very formal, formal, neutral, informal, very informal); the second specifies these traits by means of correspondences (eg very formal corresponds to formal, informal relates to jargon, etc.).
9.	Transmission channel field	This section describes language transmission (oral, written, or both).
10.	Meanings-paraphrase field	It is a virtual space where we can find the meaning of the phraseology drawn from two or more monolingual dictionaries.
11.	Field of the examples taken from the corpora	It is a section containing the meanings in context. In this section, the authors offer the opportunity to consult the most well-known corpora for each language.
12.	Thesaurus field	This is a section that shows synonyms that can be similar words or phrases.
13.	Collocation field	It is a space that offers the possibility to consult all the collocations in which the phraseological unit appears.

Tab. 1

This presentation calls for several remarks. First, we can observe that the learner or teacher is encouraged to develop a phase of discovery of phraseology that is classified and presented with its variants and in context. Then, the cleavage between the mother tongue and the foreign language is overcome by a dynamic translational space that allows the user to explore literal translations, equivalents, synonyms and different meanings using dictionaries and corpora.

We stop at the introduction of the literal translation of the frozen sequence that, according to the authors (Benigni et al., 2015: 283), is useful because it «allows the lexical constituents and mental images behind the process of construction of meaning to begin learners of the L2 ».

In addition to literal translation, theorists have paid particular attention to equivalent translation because it not only provides an understanding of the syntactic, pragmatic and functional differences between the source and

target languages, but also offers the possibility of consulting quickly the signified of the frozen expression taken into consideration.

The second part of the database provides an introduction to the field of description of the frozen unit which will be examined under several aspects as we can see below:

Description Field		
Domains or Virtual Spaces		Description
<i>Semantic description</i>	i.	<u>Argumental structure</u> Representation of the argumental structure of the actants which is completed by the analysis of their fundamental semantic features (eg "laying a rabbit" shows a structure "X does not prevent Y from its absence" where X [animated subject] and Y [animated subject]).
	ii.	<u>Lexical field</u> In approaching the onomasiological approach, this section presents the lexical field of each expression on the basis of the work of Bárdosi et al. (2003). In total, the database has 25 lexical fields.
	iii.	<u>Evaluation and connotation</u> It is a working section divided into three levels: 1) positive evaluation (approval, admiration, flattery); 2) negative evaluation (disapproval, sarcasm, contempt) and 3) neutral evaluation. In addition, the section shows the "intrinsic connotation" of the frozen unit when there are fixed sequences that appear positive but are not.
<i>Syntactical Description</i>	i.	<u>Obligatory elements in a construction</u> (eg « poser un lapin », engl. « to pose a rabbit » shows two actants X and Y, the verb form "to pose" and the object "rabbit").
	ii.	<u>Possibles transformations</u> The possibility of transforming the position of the noun for noun phraseologisms, of modifying the verbal diathesis (active, passive, reflexive), etc.
<i>Morphological Description</i>	i.	Characteristics of the constituents of the frozen sequence It is the description of modes, times and verbal voices, methods of composition and derivation and so on.
<i>Pragmatical Description</i>	i.	Container (one or more people).
	ii.	Relationships between speakers (near, distant, mixed).
	iii.	Phenomena of hierarchy (especially for Eastern languages).
	iv.	The section gives a description of speech acts (eg expression of salutation, imploration, forgiveness, etc.) and explains the intrinsic semantic connotation, the speaker's illocutionary force, the context and the discourse situation.
<i>Description of discursive elements</i>	i.	Analysis of the role of discourse markers using context and intonation.
	ii.	Historical-cultural characteristics This is a section that allows the learner or teacher to know the motivation of certain images or metaphors within the frozen sequence.
	iii.	Prosodic and gestural characteristics This is a section where the learner or teacher can find a hyperlink to audio-visual materials.

Tab. 2

The description of this second part of the project allows us to understand the importance of a holistic approach to phraseology. In fact, the authors, as we can see from this table, allow the user to have first-hand information on the characteristics of the frozen unit being considered. This way of making explicit the mechanisms and strategies of fixation is not only an advantage to compensate for the lexicographic need of translators or experts of the phraseology, but it offers above all a practical tool, on the one hand, to the teachers who can refine their educational sequences thanks to the descriptions given by the authors; on the other hand, learners who are keen to improve their foreign language proficiency level in a fast and accessible way.

The novelty of this project is, first of all, to apply the construction grammar approach that it seems very little used in the work on language didactics.

Secondly, considering phraseology as a semantic unit admits the possibility of thinking about motivation and the coexistence of several meanings.

Finally, the interlinguistic perspective favors a comparative examination of an onomasiological point of view on the differences and similarities between neighboring languages and languages belonging to different linguistic groups.

As De Knop and Mollica (2015), Mollica (2015) and Schafroth (2015) have shown, this is an approach that, although it still needs to be developed, reveals several didactic advantages as it helps to foster the recognition of fixed expressions by a guided accompaniment that tries to show the user all aspects involved in the creation of the concept of phraseology.

For adapting these theoretical approaches to French university classrooms, we have modified, mixed and enriched the models of Ettinger, Bárdosi and Lüger and the schemas proposed by the FRAME project.

As we could see below, we have proposed to the students a model where they could read an example, concerning the analysis of the French collocation “parler chiffon”, engl. “to speak rag”.

The schema is divided in three parts: 1) the expression in French language; 2) the expression in another foreign language or in mother-tongue language ; 3) description of the expression.

In the first section, we asked students to fill the following fields: 1) Variant Field; 2) Field of the examples taken from the French corpora; 3) Meanings-paraphrases field; 4) French Synonyms field.

Briefly, the first one concerns the possibility to describe the variants of the fixed expression that could be colloquial, vulgar, regional, formal, archaic or jargon, but also “very formal, formal, neutral, informal, very informal”. Students have to fill the second section using the cited online corpora to show the expression in different contexts.

In section number 3, we initiate students to use monolingual or bilingual dictionaries because we ask them to describe the meaning of the fixed expression. Last section is dedicated to synonymy and, in particular, students have to search similar words or similar fixed phrases using the synonym or monolingual dictionary.

The second part of the our phraseological entry-form concerns the fixed expression in one or two foreign languages or in the mother-tongue.

In University Sorbonne Nouvelle, the majority of students were Spanish mother-tongue studying Spanish and English; a little group speaks French as L2 and studied Spanish and English; there was a little number of French mother-tongue that studied English and Spanish or English and Russian; others were Italian studying French and English; Chinese students studying French and English; Bulgarian students studying Russian and French.

Differently, in Dokuz Eylül University, the entire classroom was Turkish mother-tongue and they studied French and German.

Our schema is divided in three parts: 1) literal translation field; 2) non-literal translation field; 3) context-translation field.

Parler chiffon		
L'expression en Langue Française		
	Champs	Description
1.	Champ des variantes	<p>Variante colloquiale et informelle : Parler chiffons (TLFi, 2018)</p> <p>Variante colloquiale et informelle : Parler toilette (TLFi, 2018)</p> <p><i>C'est une section où l'apprenant insère les variantes de l'unité quand elles existent. Les variantes peuvent être représentées par des variations minimales au niveau lexical, morphologique et syntaxique. En outre, les variantes sont organisées selon des axes diastratique, diatopique et diaphasique qui présentent les options suivantes : colloquial, vulgaire, régional, soutenu, archaïque et jargon. Il faut ajouter également le degré de formalité (très formel, formel, neutre, informel, très informel).</i></p>
2.	Champ des exemples tirés de corpus (en français)	<p>- C'est juste pour parler chiffon alors. (rss.feedsportal.com, gecrawlt am 30.04.2011)(ULC, 2018).</p> <p>- Exceptions qui confirment la règle, cinq cobayes, héros des temps modernes, ont accepté de <u>parler chiffon</u> devant l'objectif de Véronique Vial. (feedproxy.google.com, gecrawlt am 09.04.2011). (ULC, 2018).</p> <p><i>Il s'agit d'une section contenant les signifiés en contexte. Dans cette section, l'apprenant devra insérer les liens à des corpus en ligne.</i></p>
3.	Champ de la paraphrase-signifiés	<p><i>P. ext. Parler de choses frivoles, futiles. On s'arrête pour discuter avec eux, parler chiffons et mode littéraire (BRASILLACH, Pierre Corneille, 1938, p. 122) (TLFi, 2018).</i></p> <p><i>C'est un espace où les apprenants peuvent insérer les signifiés du phraséologisme tirés d'un ou plus dictionnaires monolingues (français).</i></p>
4.	Champ des synonymes français	<p>Parler de vêtements (généralement entre femmes) (Collins, 2018)</p> <p><i>C'est une section qui montre les synonymes qui peuvent être des mots ou des expressions figées similaires.</i></p>

Tab. 3

We suppose that this schema is not only useful for stimulating students to understand translation processes in case of figurative meaning, literal meaning or context meaning, but also to work on translation tools. Indeed, we ask students to fill this entry-form using a lot of lexicographic tools and this could help them to be aware and to understand concretely the sense of a “translation” through different examples.

L'expression en Langue Étrangère (deux langues étudiées ou connues)		
1.	Champ de la traduction littérale (traduction mot à mot)	<p>Traduction litt. anglaise : to speak rag (Collins en ligne, 2018)</p> <p>Traduction litt. espagnole : hablar trapo (Collins en ligne, 2018)</p> <p><i>L'apprenant insère la traduction littérale en langue étrangère (dans les deux langues étudiées ou connues) de l'unité figée française.</i></p>
2.	Champ de la traduction non littérale	<p>Traduction anglaise : talk about clothes; talk women talk (Collins en ligne, 2018)</p> <p>Traduction espagnole : hablar de trapos ; tu sabes, conversación de chicas (Collins en ligne, 2018)</p> <p><i>L'apprenant doit insérer un/deux équivalent/s de l'expression dans les deux langues connues.</i></p>
3.	Champ de la traduction en contexte	<p>ANGLAIS</p> <p>- Allez, viens grignoter un morceau et... <i>parler chiffons</i>. = Come and have a midnight snack with me and some... <u>girl talk</u> (ContextReverso, 2018).</p> <p>- Now, do you want to <u>talk about clothes</u> like a girl? = Now, do you want to <u>talk about clothes</u> like a girl? (ContextReverso, 2018).</p> <p>ESPAGNOL</p> <p>- Allez, viens grignoter un morceau et... <i>parler chiffons</i>. = Ven a medianoche a comer algo y tendremos una... <u>conversación de chicas</u>. (ContextReverso, 2018).</p> <p>- Nora Kernan, Horace et moi ne venons pas ici le jeudi pour <i>parler chiffons</i>. = Nora Kernan, Horace y yo no venimos aquí cada jueves para <u>hablar de trapos</u>. (ContextReverso, 2018).</p> <p><i>C'est un espace où les apprenants peuvent insérer les traductions en contexte dans les langues connues.</i></p>

Tab. 4

Last part of our phraseological entry-form consists in the linguistic description of the fixed expression. Specifically, we ask students to fill three sections: 1) the category of phraseology field; 2) the syntactical, semantic and pragmatical description; 3) the video-phraseologism.

First of all, we would like that students recognize the category of phraseological units (i.e. collocation, locutions, idioms, proverbs and so on) taken into consideration. They could describe the category thanks to lessons dedicated to the introduction of phraseology classification.

As we could see below, the second sections is divided in multiple sub-sections where we ask students to describe the fixed expression: 1) argumental structure using the constructions grammar approach ; 2) lexical field; 3) semantics motivation using cognitive semantics approach; 4) possible transformations; 5) origin of the expression, etymology and historical and cultural characteristics.

This kind of description could helps students to build not only a phraseological knowledge about the analyzed expressions, but also they could reflect on semantic and cultural motivation. This could give to students the opportunity to develop mental connections among different concepts, to observe linguistic relativity and to understand that people experience concepts and they structure ideas in different ways and, moreover, each cultural and social system shows a metaphorical coherence (Lakoff & Johnson, 1980).

Description de l'expression												
1.	Champ des types de phraséologisme	<p>Parler chiffon = collocation</p> <p><i>Dans cette section, l'élément phraséologique est classé comme expression idiomatique ou idiom, un proverbe, une collocation, un énoncé idiomatique, une locution verbale, une locution adjectivale, une locution nominale, une locution adverbiale.</i></p>										
2.	Description syntaxique, sémantique et pragmatique	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20%; vertical-align: top;"><u>Structure argumentale</u></td> <td style="vertical-align: top;"> <p>Parler chiffon</p> <p>X [groupe nominal GN, 1° actant, sujet animé car il s'agit d'une personne] parle [verbe monovalent] de [dans ce cas, la préposition est omise] Y [groupe prépositionnel, circonstant, objet inanimé car il s'agit de chiffon]</p> <p><i>Représentation de la structure argumentale du verbe. Il faut dire la valence verbale (le verbe peut être avalent, monovalent, bivalent, trivalent) et décrire l'expression en décrivant les actants, les circonstants, les attributs. Il s'agit des notions étudiées en Syntaxe (L2) Pour approfondir : http://www.linguistes.com/phrase/semantique.html; https://fr.wikipedia.org/wiki/Valence_(linguistique).</i></p> </td> </tr> <tr> <td style="vertical-align: top;"><u>Champ lexical</u></td> <td style="vertical-align: top;"> <p>Parler chiffon = <u>Communiquer</u> à propos des <u>vêtements</u> = champ lexical de la mode</p> <p><i>L'apprenant, en abordant l'approche onomasiologique, présente le champ lexical de chaque expression.</i></p> </td> </tr> <tr> <td style="vertical-align: top;"><u>Motivation sémantique</u></td> <td style="vertical-align: top;"> <p>Domaine source = le chiffon peut représenter un objet inutile, superflu, qu'on jette.</p> <p>Mapping = les objets inutiles rentrent dans le concept de vanité</p> <p>Domaine cible = parler des choses futiles et superficielles</p> <p><i>Il s'agit d'une section qui permet aux apprenants d'insérer la motivation sémantique de l'expression à l'aide d'une métaphore conceptuelle (Lakoff et Johnson, 1988 et 1999).</i></p> </td> </tr> <tr> <td style="vertical-align: top;"><u>Les transformations possibles</u></td> <td style="vertical-align: top;"> <p>Nous pouvons retrouver la forme à l'infinitif</p> <p>Je suis venue <i>parler chiffons</i> ; Une demi-heure en tête-à-tête à <i>parler chiffons</i> (ContextReverso, 2018).</p> <p>Nous pouvons observer également des formes conjuguées comme pour : Quand Hollande parle chiffon à l'Elysée (Bfmtv.com).</p> <p><i>L'apprenant dit s'il y a la possibilité de transformer la position du nom pour les phraséologismes nominaux, de modifier la diathèse verbale (active, passive, réflexive), etc. Cela doit être fait à l'aide de corpus ou en cherchant sur Google les formes possibles (infinitif, passif, formes conjuguées, formes interrogatives, formes négatives) qui montrent les changements.</i></p> </td> </tr> <tr> <td style="vertical-align: top;"><u>Origine de l'expression, étymologie, Caractéristiques historiques-culturelles</u></td> <td style="vertical-align: top;"> <p>Les chutes qui provenaient du travail des cousettes était des "bouts de chiffons" qui étaient utilisés pour les travaux d'entretien de l'atelier.</p> <p>De "bout de chiffon" on est passé à "chiffon".</p> <p>Un chiffonnier est celui qui ramassait les "bouts de chiffons" pour d'autres usages comme bourreler les coussins des fauteuils, confectionner d'autres vêtements etc...</p> <p>Si dans le langage courant d'aujourd'hui "chiffon" désigne communément une étoffe pour faire le ménage, "discuter chiffons" fait référence à "parler de la mode, de la tendance" (il y a même une extension : "parler chiffon" = "parler de la pluie et du beau temps" = parler de futilités".</p> </td> </tr> </table>	<u>Structure argumentale</u>	<p>Parler chiffon</p> <p>X [groupe nominal GN, 1° actant, sujet animé car il s'agit d'une personne] parle [verbe monovalent] de [dans ce cas, la préposition est omise] Y [groupe prépositionnel, circonstant, objet inanimé car il s'agit de chiffon]</p> <p><i>Représentation de la structure argumentale du verbe. Il faut dire la valence verbale (le verbe peut être avalent, monovalent, bivalent, trivalent) et décrire l'expression en décrivant les actants, les circonstants, les attributs. Il s'agit des notions étudiées en Syntaxe (L2) Pour approfondir : http://www.linguistes.com/phrase/semantique.html; https://fr.wikipedia.org/wiki/Valence_(linguistique).</i></p>	<u>Champ lexical</u>	<p>Parler chiffon = <u>Communiquer</u> à propos des <u>vêtements</u> = champ lexical de la mode</p> <p><i>L'apprenant, en abordant l'approche onomasiologique, présente le champ lexical de chaque expression.</i></p>	<u>Motivation sémantique</u>	<p>Domaine source = le chiffon peut représenter un objet inutile, superflu, qu'on jette.</p> <p>Mapping = les objets inutiles rentrent dans le concept de vanité</p> <p>Domaine cible = parler des choses futiles et superficielles</p> <p><i>Il s'agit d'une section qui permet aux apprenants d'insérer la motivation sémantique de l'expression à l'aide d'une métaphore conceptuelle (Lakoff et Johnson, 1988 et 1999).</i></p>	<u>Les transformations possibles</u>	<p>Nous pouvons retrouver la forme à l'infinitif</p> <p>Je suis venue <i>parler chiffons</i> ; Une demi-heure en tête-à-tête à <i>parler chiffons</i> (ContextReverso, 2018).</p> <p>Nous pouvons observer également des formes conjuguées comme pour : Quand Hollande parle chiffon à l'Elysée (Bfmtv.com).</p> <p><i>L'apprenant dit s'il y a la possibilité de transformer la position du nom pour les phraséologismes nominaux, de modifier la diathèse verbale (active, passive, réflexive), etc. Cela doit être fait à l'aide de corpus ou en cherchant sur Google les formes possibles (infinitif, passif, formes conjuguées, formes interrogatives, formes négatives) qui montrent les changements.</i></p>	<u>Origine de l'expression, étymologie, Caractéristiques historiques-culturelles</u>	<p>Les chutes qui provenaient du travail des cousettes était des "bouts de chiffons" qui étaient utilisés pour les travaux d'entretien de l'atelier.</p> <p>De "bout de chiffon" on est passé à "chiffon".</p> <p>Un chiffonnier est celui qui ramassait les "bouts de chiffons" pour d'autres usages comme bourreler les coussins des fauteuils, confectionner d'autres vêtements etc...</p> <p>Si dans le langage courant d'aujourd'hui "chiffon" désigne communément une étoffe pour faire le ménage, "discuter chiffons" fait référence à "parler de la mode, de la tendance" (il y a même une extension : "parler chiffon" = "parler de la pluie et du beau temps" = parler de futilités".</p>
<u>Structure argumentale</u>	<p>Parler chiffon</p> <p>X [groupe nominal GN, 1° actant, sujet animé car il s'agit d'une personne] parle [verbe monovalent] de [dans ce cas, la préposition est omise] Y [groupe prépositionnel, circonstant, objet inanimé car il s'agit de chiffon]</p> <p><i>Représentation de la structure argumentale du verbe. Il faut dire la valence verbale (le verbe peut être avalent, monovalent, bivalent, trivalent) et décrire l'expression en décrivant les actants, les circonstants, les attributs. Il s'agit des notions étudiées en Syntaxe (L2) Pour approfondir : http://www.linguistes.com/phrase/semantique.html; https://fr.wikipedia.org/wiki/Valence_(linguistique).</i></p>											
<u>Champ lexical</u>	<p>Parler chiffon = <u>Communiquer</u> à propos des <u>vêtements</u> = champ lexical de la mode</p> <p><i>L'apprenant, en abordant l'approche onomasiologique, présente le champ lexical de chaque expression.</i></p>											
<u>Motivation sémantique</u>	<p>Domaine source = le chiffon peut représenter un objet inutile, superflu, qu'on jette.</p> <p>Mapping = les objets inutiles rentrent dans le concept de vanité</p> <p>Domaine cible = parler des choses futiles et superficielles</p> <p><i>Il s'agit d'une section qui permet aux apprenants d'insérer la motivation sémantique de l'expression à l'aide d'une métaphore conceptuelle (Lakoff et Johnson, 1988 et 1999).</i></p>											
<u>Les transformations possibles</u>	<p>Nous pouvons retrouver la forme à l'infinitif</p> <p>Je suis venue <i>parler chiffons</i> ; Une demi-heure en tête-à-tête à <i>parler chiffons</i> (ContextReverso, 2018).</p> <p>Nous pouvons observer également des formes conjuguées comme pour : Quand Hollande parle chiffon à l'Elysée (Bfmtv.com).</p> <p><i>L'apprenant dit s'il y a la possibilité de transformer la position du nom pour les phraséologismes nominaux, de modifier la diathèse verbale (active, passive, réflexive), etc. Cela doit être fait à l'aide de corpus ou en cherchant sur Google les formes possibles (infinitif, passif, formes conjuguées, formes interrogatives, formes négatives) qui montrent les changements.</i></p>											
<u>Origine de l'expression, étymologie, Caractéristiques historiques-culturelles</u>	<p>Les chutes qui provenaient du travail des cousettes était des "bouts de chiffons" qui étaient utilisés pour les travaux d'entretien de l'atelier.</p> <p>De "bout de chiffon" on est passé à "chiffon".</p> <p>Un chiffonnier est celui qui ramassait les "bouts de chiffons" pour d'autres usages comme bourreler les coussins des fauteuils, confectionner d'autres vêtements etc...</p> <p>Si dans le langage courant d'aujourd'hui "chiffon" désigne communément une étoffe pour faire le ménage, "discuter chiffons" fait référence à "parler de la mode, de la tendance" (il y a même une extension : "parler chiffon" = "parler de la pluie et du beau temps" = parler de futilités".</p>											

			<p>Pour remplir cette section, l'apprenant doit chercher sur Google l'expression en tapant « origine de cette expression » ou en cherchant sur les sites Internet suivants :</p> <p>http://www.expressio.fr/, https://www.lexilogos.com/etymologie.htm, http://www.alyon.asso.fr/litterature/regles/origine_des_expressions.html, http://www.expressions-francaises.fr/annuaire-expressions-francaises.html</p>
--	--	--	---

Tab. 5

3.	<i>Vidéo-phraséologie</i>	<p>Vidéo</p> <p><i>L'apprenant doit créer une vidéo dans laquelle explique l'expression en montrant la signification littérale et la signification figurée, en suggérant l'origine de l'expression et en donnant des exemples en français ou dans une autre langue. La vidéo doit durer max 5 minutes et chaque apprenant doit faire sa propre vidéo en choisissant parmi les 5 expressions examinées par ce tableau. Je vous conseille de choisir une expression qui a un sens figuré pour avoir la possibilité de mieux expliquer les différences entre l'aspect littéral et l'aspect figuré ou métaphorique. Il faudrait faire des vidéos comme les suivantes :</i></p> <p>https://www.youtube.com/watch?v=EOFzkB4ksn8</p> <p>https://www.youtube.com/watch?v=42tfaIewoyw</p> <p>https://www.youtube.com/watch?v=eqIH-9VJmvQ</p>
----	---------------------------	---

Tab. 6

The last part of our phraseological entry-form concerns the creation of a video-phraseologism. This is based on the suggestions of Giacoma (2012 and forthcoming) that emphasized the benefits of video-based phraseology because she affirms the importance of highlighting all the pragmatic aspects that contribute to the enunciation of a frozen expression.

For our study, we ask students to make a video explaining the expression by showing the literal meaning and the figurative meaning, suggesting the origin of the expression and giving examples in different languages.

Students made different kinds of videos: some chose to do a sort of frontal lesson in which they explain the meanings, the origins and the translation in context and without context of the expressions. Others, instead, preferred to create more explicit videos in which they realize the enunciative situation concerning the expression. Finally, others have created more anonymous videos in which the slides explain to the public the characteristics of the expression. Sometimes the latter are accompanied by the voice of the student; at other times, there is only background music. All the videos are the result of the students' creativity and allowed them to have fun using linguistics and lexicography and the media means that are specific to their generation⁶⁰.

3. Results and Discussion

Even if our study needs a deeper analysis of the data and an extension of the sampling, here we limit ourselves to show the results regarding the analysis of the correct answers and the errors of the entry-forms drawn up

⁶⁰ To consult the videos, given the impossibility of inserting them in the article, we invite you to write to us or consult the YouTube page "Izmir parle français", engl. Izmir speaks French" in which many videos of the classrooms concerning the Dokuz Eylül University have been uploaded.

and compiled by the two groups of students. In order to keep the data of the two differentiated groups, we have chosen to analyze the results of each individual group separately.

Our interest is to verify how this type of teaching unit has been implemented by students by stimulating various activities such as: 1) identification of lexicalized units, 2) recognition of synonyms, 3) interlingual translation with and without context, 4) the morphological, syntactic, semantic and etymological description of the expression.

3.1 Results concerning the 1st group

The first group of students of the Sorbonne Nouvelle University is composed of 18 students who analyzed 90 frozen expressions, 5 expressions for each student⁶¹. Their level in French, as already mentioned, is equal to B2 / C1.

Most of them managed to fill in the entry-form correctly. In fact, students have understood the use of lexicographic tools such as online dictionaries and corpora and have been able to propose not only good translation examples without context and in context, but also to expand using literary and specialist contexts.

Even the exercise of the synonyms shows a greater metalinguistic awareness. Thus, for example, a student proposed as a synonym for the French expression "faire peau neuve", engl. "Change one's look" the following statement «se rénover entièrement; devenir quelque chose ou quelqu'un de nouveau», engl. «to renovate entirely; become something or someone new». Similarly, a student facing the French expression "se faire violence", engl. "Force yourself to do [sth]" proposed the following synonymic verbs fr. "Se contenir, se forcer", engl. "To contain oneself, to force oneself".

In most cases we verified the multilingual translations proposed by the students and we observed that they are correct both because the students, using online resources, were able to easily access different types of translations, and because they were aware to analyze the non-literal meaning of a frozen expression.

However, we observed some gaps in establishing a phraseological category. For example, in most cases students fail to identify the difference between a collocation and a locution such as, for example, in the case of the verbal locution "payer le prix", engl. "Pay a high price" considered by a student as a collocation.

There are other cases in which students confuse idioms with locutions such as, for example, in the case of the French expression "reprendre du poil de la bête", engl. "To take some hair again from off the beast" considered by a student a verbal locution. There are other examples that highlight students' difficulties in recognizing the phraseological type as, for example, in the case of idioms fr. "Avoir un cœur de pierre", engl. "To have a cold heart" and fr. "Poser un lapin", engl. "Stand somebody up" considered verbal locutions by two students.

Furthermore, a lot of students do not recognize the argumental structure of a frozen expression as we could see in the schema below. This depends on limiting the teaching of linguistics a few hours a week and a lack of in-depth analysis that would give students the possibility of reflecting in a metalinguistic manner on their own language and learning languages. In fact, during the semester, we tried to explicate the construction grammars

⁶¹ Below the expressions taken into consideration by the students: Papier-monnaie, Création monétaire, Taux de croissance, Taux d'intérêt réel, Honorer ses dettes, Par le biais, Faire grand cas, Nourrir les espoirs, Au cœur de, Loin s'en faut, En vogue, Contre-culture, Savant-fou, En tout cas, Au sein de, Tomber dans les pommes, Poser un lapin, Froid de canard, Tomber des nues, Etre fleur bleue, Appuyer sur le champignon, La place du mort, Conduire comme un pied, Partir sur les chapeaux de roues, Tête à queue, Avoir un cœur de pierre, Mettre les pieds dans le plats, De but en blanc, Être au septième ciel, A distance, Ouvrir ses portes, Intelligence artificielle, Réseau social, Conscience de masse, Faire le pas, Fenêtre d'opportunité, Se débarrasser de soucis, Toit pour dormir, Se tourner vers, Dans le cadre de, Se prononcer, Conduire à, vis-à-vis, En amont, Sur mesure, Au bon moment, Asseoir sa croissance, Se faire violence, Prêter à, Voir la paille dans l'œil du voisin et ne pas voir la poutre dans le sien, Endosser les traits, Payer le prix, Parler français comme une vache espagnole, Ne pas valoir un pet de lapin, Martel en tête, Gagner sa croûte, Casser du sucre sur la tête de quelqu'un, Lavage de cerveau, Voir le jour, En vue de, Suivre le scénario, Réalité objective, De plus en plus, Mettre en avant, Intelligence artificielle, Aux côtés de, A condition que/de, Se faire entendre, Affirmer son identité, Entrer en contact, Faire objet (de), Faire porter la voix, Jouer un rôle, Endosser un personnage, Endosser un rôle, Revêtir un habit, Faire bon ménage, Se donner les moyens, Relâcher notre prise, Point de vue, Prendre en compte, Faire peau neuve, Reprendre du poil de la bête, Avoir quelque chose dans le ventre, Remettre sur le tapis, Mordre la poussière, Adopter une loi, Au cœur de, Lancer un sujet, Show-biz.

theory, but students did not master linguistic concepts concerning syntax and argumental structure of a compound.

A student proposed the following argumental description for the French expression “honorer ses dettes”, engl. “honoring your debts” «X [GN nominal group, 1° actant, animated subject because it is a person] honors [monovalent verb] Y [prepositional group, circumstance, inanimate object because it is about debts]»⁶² where we could observe a misunderstanding of prepositional group or verbal valency.

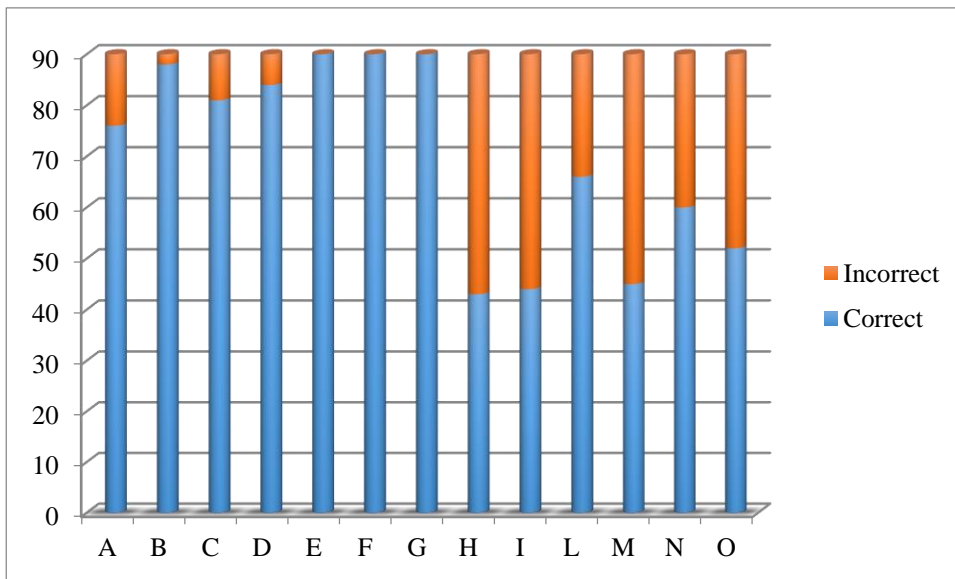
Moreover, we observe some difficulties in the identification of the semantic motivation of a fixed expression. Indeed, students show difficulties in understanding and delimiting concepts, categories and prototypes.

From this perspective, we could observe that some students could not identify the elements in the source and in the target domains. A student, for example, proposed the following description of the conceptual domains concerning the French expression “se faire entendre”, engl. “to raise your voice” «Source domain = perceive through hearing the noises, the sound; Mapping = listen, understand an idea; Target area = going out prominently, being listened to»⁶³.

A - Identification of variant field in French language
B - Identification of the examples taken from the French corpora
C - Recognition of meanings-paraphrases field in French language
D - Recognition French synonyms
E - Literal translation
F - Non-literal translation
G - Context translation
H - Classification of phraseological category
I - Recognition of argumental structure
L - Identification of the lexical field
M - Description of semantic motivation
N - Recognition of possible transformations
O - Description of the origin of the expression, etymology and historical and cultural characteristics.

⁶² Original text: «Honorer ses dettes = X [groupe nominal GN, 1° actant, sujet animé car il s’agit d’une personne] honore [verbe monovalent] Y [groupe prépositionnel, circonstant, objet inanimé car il s’agit de dettes]».

⁶³ Original text: «Domaine source = percevoir par l’ouïe les bruits, le son ; Mapping = écouter, comprendre une idée ; Domaine cible = sortir en évidence, se faire écouter».



Tab. 7

3.2 Results concerning the 2nd group

The second group of Dokuz Eylül University's students is made up by 21 students who analyzed 21 frozen expressions, one for each student⁶⁴. Their French language level is B1/B2 and the majority among them filled the entry-form quite correctly.

The ease of being able to access online dictionaries and the discovery of the use of "corpora" has been a great benefit to these students who have a level of French lower level than the group mentioned above.

Most filled the entry-form by specifying the use of the expression in context and without, proposing correct synonyms and translations in Turkish in context and without context.

We can, for example, observe that a student in translating the expression "C'est Byzance!", Engl. "It's Byzantium!" offered us a lot examples of non-literal translation models attesting that there is a metalinguistic reasoning: «Bir kuş sütü eksik (only the milk of birds is missing); Su gibi akmak (flow afloat); Cennet gibi (like paradise); Yok yok (nothing is missing); Rüya gibi (like a dream)»⁶⁵.

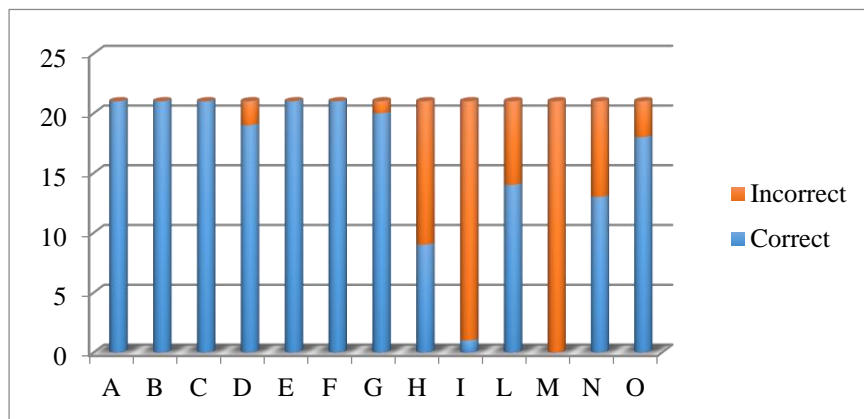
Another student proposed a perfect non literal translation for the French idiom "parler comme une vache espagnole", Engl. "speak like a Spanish cow" that in Turkish is "Fransızca'yı katletmek", Engl. "to kill French language". Another interesting non literal translation that confirms the usefulness of our didactic unit is Fr. "une armée mexicaine", Engl. "a Mexican army" that in Turkish becomes "İki kaptan bir gemiyi batırır, Engl. "Two captains sink a ship".

⁶⁴ Below the expressions taken into consideration by the students: Voir Naples et mourir, Perdre son latin, C'est Byzance!, Une réponse de Normand, Fumer comme un Turc, Travailler pour le roi de Prusse, Le fléau de Dieu, Boire comme un templier, filer à l'anglaise, parler français comme une vache espagnole, traiter quelqu'un à la turque/traiter de Turc à More, Jouer à la turque, une querelle d'Allemands, mettre la main sur son cœur à la façon des Turcs/porter la main sur son cœur à la façon des Arabes, boire un café à la turque, Une armée mexicaine, une promesse de Gascon, fort comme un Turc, lent comme un Suisse, manger en Suisse/boire/faire quelque chose en Suisse, le téléphone arabe, révérences à la turque, Un vizir aux sultans fait toujours quelque ombre.

⁶⁵ Original text: «Bir kuş sütü eksik (il ne manque que du lait d'oiseau); Su gibi akmak (couler à flot); Cennet gibi (comme le paradis); Yok yok (il ne manque rien); Rüya gibi (comme un rêve)».

As previously mentioned, the students of the second group, having never had an explicit teaching of linguistics, showed some difficulties and gaps in understanding the categorical differences in phraseology and in recognizing the argumentative structure as we could see in the schema below.

Although the identification of the lexical field is quite satisfactory, these students also showed difficulties in understanding the conceptual domains of frozen expressions.



A - Identification of variant field in French language
B - Identification of the examples taken from the French corpora
C - Recognition of meanings-paraphrases field in French language
D - Recognition French synonyms
E - Literal translation
F - Non-literal translation
G - Context translation
H - Classification of phraseological category
I - Recognition of argumental structure
L - Identification of the lexical field
M - Description of semantic motivation
N - Recognition of possible transformations
O - Description of the origin of the expression, etymology and historical and cultural characteristics.

Tab. 8

4. Conclusion

Our study has tried to develop a didactic activity aimed at creating an "active" knowledge of the word. Students, in fact, discovering the use of dictionaries and corpora, can verify the meaning without context and in the context of an expression, they can understand the difference between a literal and non-literal translation, they experiment linguistic concepts that they had not previously faced. The entry-form is an educational experience that furnishes and provides elements to expand not only the vocabulary and terminology of students, but also to encourage students to study in depth the "other" language through differences and similarities with the mother tongue or with the other studied languages.

From this form of approach, we understand that the learner can gain more autonomy because he is forced to look for constructions in a comparative way making incursions in the cultural-historical context of his mother-tongue language and the foreign language and he/she can thus dynamically reconstruct the conceptual networks at the base of a frozen expression; he/she can also access the meaning of the expression by a reflection on its origin.

Furthermore, reflecting on the conceptual domains that are inherent in the expressions, allows the student to access a meta-cognitive reflection on the categorization process. This helps the student to approach not only the universal mechanisms involved in the construction of language, but also a cultural and phenomenological relativism.

5. References

- Benigni, V.; Cotta Ramusino, P.; Mollica, F. & Schafroth, E. (2015). «How to Apply CxG to Phraseology: A Multilingual Research Project». In *Journal of Social Sciences*. DOI: 10.3844/jsssp.2015.
- Boers, F. ; Deconinck, J. & Lindstromberg, S. (2010). « Choosing motivated chunks for teaching ». In De Knop, S. ; Boers, F. & De Rycker, A. (Eds). *Fostering Language Teaching Efficiency through Cognitive Linguistics*. Berlin/New York : De Gruyter Mouton, 239-258.
- Cacciari, C. & Tabossi, P. (1988). « The Comprehension of Idioms », *Journal of Memory and Language* 2, 668-683.
- Cavalla, C., & Labre, V. (2009). «L'enseignement en FLE de la phraséologie du lexique des affects». In A. Tutin & NovakovaI. (Eds), *Le lexique des émotions et sa combinatoire lexicale et syntaxique*. Grenoble: Ellug, 297-316.
- Chi, A. (2003). *An empirical study of the efficacy of integrating the teaching of dictionary use into a tertiary English curriculum in Hong Kong*. Hong Kong: Language Centre Hong Kong University of Science and Technology.
- Chi, A. (2011). «When dictionaries support vocabulary learning, where to begin?». In K. Akasu & S. Uchida (Eds) *ASIALEX '11 Kyoto Proceedings: Lexicography: Theoretical and Practical Perspectives*, 76–85.
- Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
- Dobrovol'skij, D. (1995). *Kognitive Aspekte der Idiom Semantik: Studien zum Thesaurus deutscher Idiome*. Tübingen: Gunter Narr.
- Dobrovol'skij, D. (1997). *Idiome im mentalen Lexikon: Ziele und Methoden der kognitivbasierten Phraseologieforschung*. Trier: Wissenschaftlicher Verlag Trier.
- Dobrovol'skij, D. (2004). «Idiome aus kognitiver Sicht». In Steyer, K. (Eds), *Wortverbindungen - mehr oder weniger fest*. Berlin: Walter de Gruyter, 117-143.
- Dobrovol'skij, D. O. & Piirainen, E. (2005). *Figurative Language: Cross-cultural and Cross-linguistic Perspective*. Amsterdam: Elsevier.
- Fillmore Ch., Kay P. & O'Connor P. (1988). «Regularity and idiomaticity in grammatical constructions : the case of let alone». In *Language* 64, 501-38.
- Gentner, D. (1983). «Structure-mapping: A theoretical framework for analogy». In *Cognitive Science*, 7/2, 155-170.
- Gentner D.; Holyoak, K. J. & Kokinov, B. N. (2001). *The analogical mind. Perspectives from cognitive science*. Cambridge & London: The MIT Press.
- González Rey, M. I. (2010). « La phraséodidactique en action: les expressions figées comme objet d'enseignement ». In *La Clé des Langues*. Lyon: ENS LSH/DGESCO.
- Gibbs, R. W. Jr. (1986). «Skating on thin ice: Literal meaning and understanding idioms in conversation». In *Discourse Processes*, 9/1, 17-30.
- Granger, S. & Meunier, F. (Eds) (2008). *Phraseology: an interdisciplinary perspective*, Amsterdam/Philadelphia: John Benjamins.
- Hofstadter, D. (1995). *Fluid Concepts and Creative Analogies. Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

- Jantunen, J. H. (2016). «Corpora, phraseology and dictionaries : How does corpus research intersect language teaching and learning?». In B. S. Vilas (Eds), *Collocations cross-linguistically : Corpora, dictionaries and language teaching*. Uusfilologinen Yhdistys, 97-119.
- Lakoff G. & Johnson M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.
- Lakoff G. & Johnson M. (1999). *Philosophy in the flesh – The Embodied Mind and its Challenge to Western Thought*. New-York : Basic Books.
- Langacker, R. (1987). *Foundations of cognitive grammar, Volume 1 : Theoretical prerequisites*. Stanford : Stanford University Press.
- Lewis, M. (2000). *Teaching collocation : Further developments in the lexical approach*. Hove: Language teaching publications LTP.
- Paquot, M. (in press). «Lexicography and phraseology». In B. Douglas & R. Randi (Eds), *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.
- Pawley, A., & Syder, F. H. (2000). «The one-clause-at-a-time hypothesis». In H. Riggenbach (Eds), *Perspectives on fluency*. Michigan: University of Michigan Press, 163-199.
- Prinsloo, D.J., Bothma, T.J.D., Heid, U. & Prinsloo, Daniel. J. (2017). «Direct user guidance in e-dictionaries for text production and text reception - the verbal relative in Sepedi as a case study». *Lexikos* 2017, 403-426.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1998). «The lexical item». In E. Weigand (Eds), *Contrastive Lexical Semantics*. Amsterdam: John Benjamins.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard: Harvard University Press.
- Waara, R. (2004). «Construal, Convention, and Constructions in L2 Speech». In Achard, M. & Niemeier, S. (Eds), *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*. Berlin/New York : Mouton de Gruyter, 51-76.
- Wood, D. (2006). «Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency». In *The Canadian Modern Language Review*, 63/1, 13–33.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Dictionaries and corpora

ORTOLANG, Outils et Ressources pour un Traitement Optimisé de la LANGue, <https://www.ortolang.fr/> & <http://cnrtl.fr/>, 2015.

TLFi, *Le trésor de la langue française informatisé*, <http://atilf.atilf.fr/>, 2017.

UNIVERSITÄT LEIPZIG, *Wortschatz Universität Leipzig, Korpus French*, http://corpora.uni-leipzig.de/de?corpusId=fra_mixed_2012, 2012.

FINNISH–TURKISH DICTIONARIES: PRESENT STATE AND FUTURE PERSPECTIVES

Mats-Peter Sundström

European Parliament

Swedish translation unit

Abstract

This paper purports to present an overview of the situation for bilingual lexicography between the Finnish and the Turkish languages. Admittedly, these languages may appear a somewhat unlikely language pair for dictionaries to cover. On close inspection, however, a different picture appears. According to a long-held, although currently for all practical purposes abandoned linguistic theory, Finnish and Turkish were related languages, and this notion still lives on among large sections of the general public. Thus, there may be grounds for assuming there might be vast body of dictionaries involving these two languages. Although this is manifestly not the case, a survey will be made of the situation for Finnish–Turkish–Finnish dictionaries today, focusing on the following parameters: 1) historical development 2) sizes of current dictionaries and 3) distribution between general dictionaries and LSP dictionaries 4) printed vs electronic dictionaries and finally 5) some notes on lexicography between Finnish and other Turkic languages. Subsequently will be presented some comments on potential users of these dictionaries and some general remarks on the lexical and grammatical structure of the two languages involved, although practical constraints will obliged these comments to be made almost solely from a Finnish point of view. In fact Finnish users will most likely vastly outnumber Turkish users, in so far as the number of Turkish-language persons in Finland is very limited (as opposed to in neighbouring Sweden) and Turkey is also a very popular tourist destination for citizens of Finland. The paper will end in a plea for increased cooperation between lexicographers in Finland and Turkey.

Key Words: Bilingual dictionaries, language pair: Finnish and Turkish

Introduction

At the Fourth Asialex conference in Singapore 2005 the following was pointed out: "A casual observer of the lexicographic scene may well be pardoned for concluding that lexicography, either bilingual or monolingual, is somehow bound to involve the English language" (Sundström, 2005). Indeed this feature of the lexicographic landscape has changed little since then. After all, whether on-line or in paper format, bilingual dictionaries tend largely to concentrate on major languages. The vast majority of them tend to have a major language (or a language of large usership) for both source and target language (e.g. a Chinese–English dictionary) or at least involve a major language in either one of those roles (like a Finnish–English, or a French–Polish dictionary). Indeed: "bilingual dictionaries where both the source and target language are lesser-used languages are the odd man out among bilingual dictionaries" (Sundström, 2005), and larger-size dictionaries of that kind are almost unknown. A striking example of the contrary is provided by the two-volume *Suuri suomi-ruotsi sanakirja* (Finnish–Swedish General Dictionary, Cantell et. al. 1997), published in 1997 and containing some 130 000 headwords. This is exceptional, as it is addressed to language communities of some ten million users, as is the case with Swedish and some five million users, as is the case with Finnish. The reason behind this somewhat surprising situation can basically be found in Finland's being a bilingual nation with Finnish and Swedish constitutionally recognised as national languages

Even so, it remains a solid fact that bilingual dictionaries between minor languages are few and far between. At this point, it should be strongly underlined that the term "minor language" most assuredly implies no value judgement. For the purposes of this paper, by a minor language is meant a language whose distribution is basically restricted to the national rather than the international level and whose user base in purely demographic terms often (but certainly not always) is restricted. With this definition in mind it is considered justified to apply the notion of minor language also to Turkish, although, by comparison with the Finnish language, the Turkish first-language speakers outnumber those with Finnish for their mother tongue by more than sixteen to one! And of course, if we include not only the Turkish but the Turkic languages, the practice of calling Turkish a minor language is in even greater need of a justification.

This paper proposes to study the situation with regard to bilingual dictionaries between Finnish and Turkish. This approach may give rise to the question why at all there should be any such dictionaries, given that the geographical areas where the two relevant languages are spoken are not exactly next-door neighbours. The answer essentially lies in the Ural[o]-Altaic language hypothesis, claiming an affinity between languages such as (among others) Finnish, Mongolian and Turkish. Proposed originally in the early 18th century by, among others, German philosopher Leibniz, this theory found one of its main advocates in the Finnish linguist Mathias Alexander Castrén (1813–1853). Although nowadays largely abandoned, this hypothesis has left a lasting impact on the popular mind, and is not altogether extinct among scholars, either. Suffice it to quote an Asialex conference paper from 2016, where it is pointed out that the "Turkish language forms the Altai[c] branch of [the] Ural Altaic Language Family together with Mongolian and Tungusic" (Gürlek 2016, 36). Likewise, what is probably the world's first Finnish–Turkish–Finnish dictionary said in its Finnish-language foreword: "Turkkilaiset kielet muodostavat melko laajan yhtenäisen alueen Balkanista ja Anatoliasta aina Itä-Siperiaan saakka", i.e. the Turkic languages form a fairly large uniform area from the Balkans and Anatolia all the way to East Siberia (Ekim 1983, 35).

This paper will proceed to study whether this perception has manifested itself in an interest in lexicography between Finnish and Turkish. To do so, the dictionary situation will be reviewed from two angles: historically and taking today's situation into account and also by surveying the availability of general dictionaries vs LSP dictionaries involving both the languages concerned.

Method

With respect to method, this paper may be brief and concise indeed. After relatively extensive on-line searches, fact-finding contacts were made with the following entities: 1. the publishing house WSOY, Finland's largest publisher of bilingual dictionaries. 2. Finland's national library in Helsinki, formerly known as the Helsinki University Library. 3. the Helsinki city library and 4. the Akateeminen Kirjakauppa/Akademiska Bokhandeln, the largest bookstore in Finland.

For the purposes of this paper, two lexicographic works were selected for closer scrutiny: one general language dictionary and one LSP glossary. The selection was dictated entirely by constraints of availability and accessibility. As will be pointed out further on in greater detail there is an astonishing scarcity of dictionaries between Finnish and Turkish and none of them were available in my country of stationing, i.e. Luxembourg. This being so, research had to be concentrated to the author's brief periods of homeland stay. And the search and research process was in no small measure complicated by perfectly extralinguistic factors, bearing no point in common with lexicography. Although there are at least sixteen times more people living in Istanbul than in the Helsinki metropolitan region, the latter spreads out over a geographical area almost the same size as that of Istanbul (urban sprawl, if ever there was!). Given these realities, it is obvious that the fact of the Finnish–Turkish–Finnish dictionaries being distributed mainly over libraries in Helsinki's peripheral districts, rather than in the central libraries (for reasons unknown to the author) did little to help researches.

Then, owing to the extremely high workload at the author's workplace, the European Parliament language service, right in the middle of the run-up to the European elections in late May this year, no visit to Istanbul

was possible as a complement to the investigations undertaken in the author's native Finland. Regrettable though this situation was, there was no amending it, much to the detriment of this paper's completeness and versatility.

In the general dictionary samples were taken covering every tenth page of the Finnish–Turkish lemma list, and likewise for the Turkish–Finnish lemma list. The glossary referred to, with Finnish for its source language and Turkish for its target language was examined *in extenso*.

Results

If the method used lent itself to short and summary description, so will the results. Should there be a need for adjectives to qualify them with, "dismal, disappointing and disastrous" would fit beautifully as a general characterization of them. First of all from a historical point of view: the earliest Finnish–Turkish dictionary identified dates back only to 1975. Strikingly enough, it was not published in Finland, but in Turkey, under the name of Suomi-Turkki Matkailusanasto – Fince-Türkçe Gezi Sözlüğü and edited by Ismael Ekim, reprinted in 1983. This work, interestingly enough, contains some 7 000 Finnish headwords with Turkish equivalents, but only 4 000 Turkish headwords with Finnish equivalents. In 1992, Finland followed suit, insofar as the Suomi–Turkki–Suomi sanakirja appeared, published by the Finnish publishing house Gummerus. One year after the Suomi–turkki-suomi taskusanakirja (Finnish-Turkish–Finnish Pocket Dictionary) appeared, edited by Jorma Atilla, a Finn, who for some time had been part of the faculty at Ardahan University, Turkey. Published by the aforementioned WSOY company this medium-sized dictionary of some 31 000 words and phrases was updated and revised in 2002, but as of March 2019, it was long since out of print and not readily available at libraries either. Thus, as to the present-day dictionary repertoire: there is only one general language dictionary on offer to the broader Finnish public today, namely the Suomi—Turkki–Suomi sanakirja, a product of the Finnish publishing house Gummerus.

Finally, as regards LSP dictionaries between Finnish and Turkish: there is but one (!) of them, or rather a glossary. Supplied with a foreword in three languages (Finnish, English and Turkish, in that order), in English it bears the title of Finnish–Turkish Glossary of Criminal Justice, and was published on-line as a thesis at the Helsinki-based Diaconia University of Applied Sciences in the autumn of 2017 by Tuula Ulug. In the English-language part of its foreword (called Abstract), this publication concludes laconically: "Because of a lack of Finnish–Turkish dictionaries, especially in specific fields, there was an urge for this kind of glossary" (Ulug 2017, p. 39). Few if any would feel inclined to dispute this opinion!

At this stage a pressing need for qualifying this unmitigatedly bleak picture of bilingual lexicography involving Finnish and Turkish makes itself felt. It will be met by a statement to the effect that what has been presented so far refers only to print format dictionaries. In the on-line world, the situation takes on a somewhat rosier complexion. Mention must be made above all of the Collins online application offering bidirectional translations between Finnish and Turkish, with some ten thousand headwords all in all. The fact however remains that a classical printed dictionary between the two languages concerned remains indeed a lone wolf and orphan in the world of lexicography.

A point in between: might there be dictionaries involving Finnish and other Turkic languages than Turkish? To the best of the author's knowledge there are none. Then, an interesting lexicographic project reportedly launched during the last years of the Soviet Union, aimed at creating dictionary links between the main languages of all the then fifteen USSR republics might have delivered dictionaries between on the one hand Estonian (a cognate language of Finnish) and Turkic languages such as Azeri, Uzbek, Kazakh, to mention but a few. A Lithuanian-origin linguist colleague of mine at the European Parliament, however, decidedly concluded that this dictionary project never materialised (cf Lillieholm 2019).

Discussion

The paper will now proceed to discussing more in detail two of the dictionaries presented above. Initially the focus will be on Suomi–Turkki–Suomi, by Heinrich Bremer and Mersa Luukkonen. Subsequently attention will be drawn to Ulug’s glossary of legal terms. Obviously, constraints of space will necessitate a concentration on some of the most basic features of these dictionaries, intended to serve as a starting point for subsequent proposals concerning the future of Finnish–Turkish–Finnish lexicography.

Focus will now be on the most readily available (in bookstores and libraries) Finnish–Turkish dictionary in Finland these days. As already mentioned, this dictionary appeared for the first time in 1992, editors: Heinrich Bremer and Mersa Luukkonen. In 2008 a revised version of it was published, this time edited by Heinrich Bremer and Mikko Virtanen and the publisher’s dictionary editing department.

The dictionary in question forms part of the publisher’s series of bilingual dictionaries called *Matkalle mukaan - matkasanakirjat*, or in English translation ”Take along while out travelling – traveller’s dictionaries”. As the name of the series suggests, these dictionaries are basically geared to cater for the tourist’s lexicographic needs when visiting an area where the dictionary language that is not his mother tongue is spoken. In this case, the main target group is clearly Finnish tourists to Turkey. As it is clearly stated in the front matter, the dictionary is ”laadittu helpoksi kielioppaaksi kaikille suomalaisille matkailijoille” i.e. prepared as an easy-to-use language guide for all Finnish travellers. This approach is reflected in the dictionary framing structure. As defined by Lehmann, this structure ”comprises a set of main sections that correspond to the chapters of a book” (Lehmann 2017).

Apart from two lemma lists (one Finnish–Turkish, the other Turkish–Finnish, each with some 4 500 headwords, there are three additional sections in the framing structure. They are as follows: a 43 page phrasebook, with phrases grouped under various headings such as (in English) ”at the hotel”, ”shopping”, ”travelling and tourism”, to mention only a few. Then follows a semi-encyclopedic section, whose heading would translate into English as ”Republic of Turkey. Basic Information for Tourists. Featured in this section are descriptions of geography, history ... and, occupying a full fourth of the section, information about local food and drink. Finally come fifteen pages: ”Short Turkish Grammar”.

It goes without saying that the level of lexicographic ambition in a bilingual dictionary of this kind is bound to be modest indeed. Even the dictionary microstructure amply reflects this fact, being cut back to its barest minimum. From a sampling it turned out that more than ninety per cent of the articles were modelled on the pattern of headword followed by one single target language equivalent, with neither labels nor explanatory metalanguage used, except in the rarest of cases, and illustrative phrases not occurring at all. Only two LSP-indicating labels are used: *anat.* for *anatomia* (no translation into English required) and ”lääket.”, equalling the Finnish noun *lääketiede* meaning medicine. What little explanatory metalanguage occurs is invariably in Finnish, and is used to disambiguate polysemous Finnish words in relation to their Turkish equivalents. A case in point is the Finnish noun *maali*, with the basic meanings of ”target” and ”goal” in English (the homonymous Finnish *maali* meaning ”paint” is not included in the lemma list), where five Turkish equivalents are supplied, all with a Finnish explanatory metalanguage item within brackets.

Sometimes, the microstructure is anything but helpful to the user. Mention may be made of the Finnish noun *siika* referring to a particular kind of fish, in English mostly called common whitefish (in Latin *Coregonus lavaretus*). For an equivalent, the dictionary offers only two Finnish words ”Turkissa tuntematon”, a phrase that means ”[this fish is] unknown in Turkey” (!).

Once the dictionary was caught out committing a factual error. For some reason, a name of a flowering plant *kruunuvuokko* was introduced, unknown in Finland but probably quite prevalent in areas of Turkey traditionally visited by Finnish tourists. About this plant, Wikipedia offers the following information: ”*Anemone coronaria*, the poppy anemone, Spanish marigold, or windflower, is a species of flowering plant in the genus *Anemone*, native to the Mediterranean region”. However, the dictionary’s Turkish equivalent for the

relevant flower is *Manisa lalesi*, which in turn in Latin is called *Tulipa orphanidea*, thus a totally different species.

The occasionally somewhat amateurish approach to headword presentation in the above-mentioned dictionary is to some degree also reflected in confusion between micro- and macrostructure. To illustrate: the Finnish headword *koko* comes with two numbered senses within the same article (1) meaning *koko* as a noun in the sense of *size* then (2) meaning *koko* as an indefinite pronoun in the sense of *all, everything*. Then, the Finnish noun *korko* is divided over two articles: one covering *korko* in the sense of "heel [of a shoe]", the other covering *korko* in the sense of "interest [rate/percentage]". As is evidenced by Lehmann, the latter technique, rather than the former, corresponds to established lexicographic practice: "Homonyms are, of course, separate entries distinguished by numbers" (Lehmann 2019).

In sum then: while a possibly handy language guide for Finnish tourists in Turkey, the dictionary discussed above meets only very modest requirements for bilingual dictionaries. This said, the discussing will proceed to deal with another representative of Finnish–Turkish bilingual lexicography, with vastly different characteristics, namely the Finnish–Turkish Glossary of Criminal Justice.

As the very title suggests, this is most eminently an LSP lexicographic product. Right from the outset, it also demarcates itself quite neatly from the dictionary discussed above. For one thing, this glossary is clearly an academic product. It was presented on-line in the autumn of 2017 as a thesis under a degree programme in community interpreting. In the front matter, also presented in English, the target group is distinctly delimited: "The glossary is meant [for] court interpreters and interpretation students" (Ulug 2017, 3), along with an explanatory statement as to why it was prepared in the first place: "Because of a lack of Finnish–Turkish dictionaries, especially in specific fields, there was an urge for this dictionary". Then, the article microstructure is succinctly and accurately explained, stating that the glossary "has been compiled in a way that the terms in the Word Table come first in Finnish. The definition is situated under the term and beneath the definition, there are eventual sentences where the term in question is used. Beside the Finnish term or sentence, one can find the equivalent term, the definition of it and a translation in Turkish" (Ulug 2017, 3). At this point, mention should be made of the fact that not only are the headwords themselves (obviously!) provided with a Turkish equivalent, but so are also the both the definitions and the usage illustrative phrases.

The academic character of Ulug's glossary is also reflected in the explicit mention of its directionality. As the author herself points out: "The glossary of this thesis helps [Finnish] interpreters and interpretation students in their work and studies. It may also help the Turkish speakers to understand criminal justice terms in Finnish if they do not want to use an interpreter.

The glossary briefly presented above is evidently a pioneering and meritorious work within the extremely restricted framework of Finnish–Turkish bilingual lexicography, and, as was already mentioned, the first of its kind focusing on LSP language. It may only be sincerely hoped and wished for that it will find successors, targeting other LSP areas involving the two language concerned.

As a final point concerning LSP lexicography involving the Finnish and the Turkish languages, mention should be made of IATE, which means Interactive Terminology for Europe. This is the European Union's (EU) term base, offering direct search pathways between all the twenty-four official EU languages and sometimes even Latin. This tool comes with interesting implications relating to the field of international affairs. Bearing this in mind: what is said hereafter must in no way be construed as in any way hypothesising, neither about Turkey's future relations with the EU, nor about whatever possible structure they might assume. Suffice it to say that a hypothetical Turkish membership would imply participation of the Turkish language in this term base, thus enabling searches between more than a million Turkish entries and their Finnish equivalents and vice versa (as of mid-April 2019, the IATE contained some 1 020 000 entries).

At this point, the discussion will turn onto its second main theme, namely the reasons for this paucity of bilingual dictionaries between Finnish and Turkish. In a way, this indirect question may be stood on its head

so as to read, initially in this paper: "Why should there be such dictionaries in the first place, given that both Finnish and Turkish are (from an international point of view) decidedly minor languages? Again, the answer may be found in the Uralo-Altaic language hypothesis as a possible factor giving impetus to interest in each other's language on the part of those having different languages from this (presumed) language family as their respective mother tongues. However, this hypothesis is not borne out by reality. The underlying reasons may conceivably be summed up as follows:

1. Bilingual lexicography involving Finnish is of very recent date in Finland also. A better understanding of this requires some information about language conditions in Finland. Most important of all: it was only in the latter part of the nineteenth century that the Finnish language was recognised as a vehicle for culture, science and public life, on a par with Swedish. Finland, as it were, only then opened up to the outside world in the language spoken by the majority of its population. Often referred to as the Finnish national revival this complex process among many other things involved the compilation of dictionaries between Finnish and the major European languages. In this context, Turkish was not considered sufficiently important, not least because of the strong position of French in the then Ottoman Empire that made a mastery of Turkish almost redundant for contacts with the country.

2. National bias. At this point the author is keenly aware of venturing onto thin ice indeed, owing to the extremely sensitive nature of the matter, but there is no shunning away from it: a lack of interest in Turkish, owing to a misguided sense of Occidental supremacy. By and large Finland's historical attitude towards Turkey and things Turkish used to coincide with that of the Occident in general, and thus, even at its most charitable, must be qualified as dismissive and aloof. Indeed, language based affinity between Finland and Turkey as a source of a positive general attitude has been very much a one-way street. In favour of Turkey, it should be added. An illustration will follow: When then newly independent Finland on February 21 1918 sent a delegation to the Ottoman empire seeking diplomatic recognition for Finland, one of the delegation members afterwards wrote: "The hospitality and friendliness [encountered in Turkey] were fabulous. Everyone, from the sultan himself and down to the journalists treated us Finns as the dear northern relatives of the Turkish people, who after thousands of years of separation had come together again" (Finnish Ankara embassy 2017, p. 2).

Needless to say, there is no way of proving a direct link between this mind-set and a lack of Finnish interest in the Turkish language and consequently in dictionary editing between Finnish and Turkish. But could not, conceivably, the effusively positive attitude towards Finland and things Finnish have spurred a Turkish lexicographic interest in the Finnish language? Possibly so, but it has to be recalled that purely material constraints, i.e. lack of resources, combined with a host of other pressing concerns, not least in the field of linguistics, probably precluded Turkish academia and scholars to engage in such pursuits.

3. Lack of a large Turkish immigrant community in Finland. Unlike neighbouring Sweden, where a massive influx of Turkish migrants started making itself felt even in the 1960's, resulting in the presence of some eighty thousand persons of Turkish origin in Sweden today, the corresponding number is ten times lower in Finland (slightly over seven thousand, according to statistics dating back to 2016). Against this background it is hardly a wonder that there is a major dictionary between Turkish and Swedish (*Turkisk-svensk, svensk-turkisk ordbok*), by Musa Güner, totalling, in its fourth edition 82 000 headwords and phrases.

Conclusion

In a certain way, the conclusions of this paper have already come to the fore in the previous sections. Summing them up again they would run as follows: The situation with regard to bilingual dictionaries involving Finnish and Turkish is, by way of putting things leniently, anything but satisfying. From this follows logically that there is an urgent need for remedial action. After all, as was recently pointed out by no lesser a body than the UNGA, United Nations General Assembly in a resolution on the role of translators, that translations are indispensable to "preserving clarity, a positive climate and productiveness in international public discourse

and interpersonal communication” (UNGA 2017, s. 2.). Now, if the cavalier and casual choice of expression may be pardoned, it is something of a perfect no-brainer that dictionaries rank among a translator’s most important tools. Evidently there are ways of overcoming a lack of direct lexicographic links between a pair of languages such as Finnish and Turkish, relying on any one of the plentiful Finnish–English and English–Turkish (and vice versa) dictionaries that are available. But a direct link will almost invariably turn out more effective and efficient for translation purposes than the method of using IMBL’s (= Intermediate bilingual dictionaries).

Though alas: dictionaries are essentially commercial ventures. This being so: would there be a market large enough to warrant the launching of a major Finnish–Turkish dictionary project of the same size as, say, the Swedish–Turkish dictionary referred to above? Probably not. Or then again: might there not, after all? In fact, the Vietnamese language community in Finland numbers only some five thousand members (as opposed to, as stated above, some seven thousand Turkish speakers). Even so, the Finnish and Vietnamese languages have enjoyed very extensive bidirectional coverage for close to twenty years with the "Suomi–Vietnam sanakirja" ['Finnish–Vietnamese dictionary'], published in 1997 by Laurent Tran-Nguyen and followed in 2002 by the companion volume "Vietnamilais-suomalainen yleissanakirja" ['Vietnamese–Finnish general dictionary'] by the same author. Both dictionaries contain over sixty thousand words. For the sake of completeness it should also be pointed out that 2018 saw the publishing of a large Finnish–Japanese–Finnish dictionary with some 48 000 headwords, at a time when the number of Japanese living in Finland and Finns studying Japanese (not to mention Japanese studying Finnish!) is limited indeed (cf Karppinen, 2018).

At this stage, however, reference must be made to Kwary’s observation: ”In creating a dictionary, lexicographers need to establish the profile of the dictionary users. Without establishing its specific users, a dictionary will simply become a display without anyone benefitting from it.” (Kwary 2018, 106). Consequently, it should be made clear from the very beginning that a major dictionary work between the Finnish and the Turkish language would from the outset, owing to practical circumstances, be primarily directed towards people taking an academic interest in either one of the two languages, with all the concomitant restrictions for market size this will imply.

Thus, here, if anywhere, a PPC, Public Private Partnership between publishers, academia, language research centres etc. in both (yes: underlining the word ”both”) Finland and Turkey might well prove crucial to making a project of this kind come true. As was pointed out regarding another pair of language, unlikely to occur in a bilingual dictionary, namely Finnish and Swahili: ”Uhusiano, an umbrella organisation for Finnish development co-operation organisations in the Morogoro area in Tanzania launched the [Finnish–Swahili] dictionary project. The organisation acquired project funding from the Department of International Co-operation of [Finland’s] Ministry for Foreign Affairs and the Department for Education and Science Policy of [Finland’s] Ministry of Education” (Ollikainen, 2002).

And on a plea for that very kind of project to be launched I should like to conclude this paper, sincerely harbouring a hope to the effect that if ever it materialises, then the Polish phrase of *turecki kazanie*, that is ”a Turkish sermon” would at least to Finns no longer be a byword for something incomprehensible, but rather an expression of a culture both possible and worthy to partake of.

References

- Atilla, J. (1995). *Suomi-turkki–suomi sanakirja*. Helsinki: WSOY.
- Bremer, H., Virtanen, M. (1992). *Suomi turkki suomi sanakirja. Matkalle mukaan*. Porvoo: Gummerus.
- Cantell, I., Martola, N., Romppanen, B., Sundström, M.-P. (1997). *Suuri suomi–ruotsi-sanakirja. Stora finsk–svenska ordboken*. Helsinki: Werner Söderström Osakeyhtiö & Kotimaisten kielten tutkimuskeskus/Forskningscentralen för de inhemska språken.
- Ekim, I. (1983): *Suomi–Turkki, Turkki–Suomi Sanakirja*. Helsinki: Printhouse.

- Finnish Embassy(2017). *Maatiedosto Turkki: Kahdenväliset suhteet*. Retrieved from <http://www.finland.org.tr/public/default.aspx?nodeid=48798&contentlan=1&culture=fi-FI>
- Güner, M. (2011). *Turkisk-Svensk Svensk-Turkisk Ordbok*. Lund: Musa Güner Förlag.
- Gürlek, M. (2016). *Turkish Lexicography: From Ottoman lexicography to Turkish dictionaries*. In: Proceedings of the 10th International Conference of the Asian Association for Lexicography. Manila: Asialex 2016.
- IATE. Interactive Terminology for Europe, consulted May 7 2019.
- Kwary, D. (2018). *The variables for drawing up the profile of dictionary users*. In: Lexicography. Journal of Asialex. Volume 4 Number 2 September 2018. Heidelberg:Springer-Verlag GmbH.
- Karppinen, T. (2018) *Suomi–japani–suomi sanakirja*. Tallinn: Arthouse.
- Laurent, Tran-Nguyen (2014): *Suomi–Vietnam sanakirja*. Helsinki: Gaudeamus.
- Lehmann, C. (2017). *Lexicography*. Retrieved from https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html
- Liliehalm, I. (2019). Verbal communication to author Apr. 15 2019.
- Ollikainen, T. (2002). *Say it in Swahili!* In Universitas Helsingiensis, retrieved from <http://www.helsinki.fi/uh/2-2002/juttu14.shtml>
- Sundström, M.P. (2005). *Intermediary languages in bilingual lexicography for lesser-used languages: the pros and cons*. In: Proceedings of the 4th International Conference of the Asian Association for Lexicography. Singapore: Asialex 2005.
- Ulug, T. (2017). *Rikosoikeuden yleissanasto suomi–turkki. Finnish–Turkish Glossary of Criminal Justice*. Helsinki: Diakonia-ammattikoulu/Diaconia University of Applied Sciences.
- UNGA (2017). *The role of professional translation in connecting nations and fostering peace, understanding and development*. Resolution 71/288 adopted by the United Nations General Assembly on 24 May 2017. New York.
- Wikipedia (2019), consulted May 7 2019.

ON THE FUTURE OF LEXICOGRAPHY AND DICTIONARIES IN TSHIVENḌA

Munzhedzi James Mafela

University of South Africa

Abstract

Tshivendḍa is one of the eleven official languages of the Republic of South Africa. It was accorded the status of official language together with the other eight indigenous African languages after the democratic elections of 1994. Afrikaans and English were the official languages long before 1994. The democratically elected government established lexicography units for all the official languages, partly to help the indigenous African languages to produce dictionaries. These languages were marginalised by the previous governments, and did not receive funding for dictionary making. Since the introduction of national lexicography units, the number of dictionaries and lexicography work increased in the indigenous African languages. However, these units face challenges because the government's grant is not enough to sustain them. Unlike Afrikaans and English national lexicography units, the indigenous African languages units cannot raise funds to sustain themselves; they depend on government funding. As a result, some lexicographers are leaving the dictionary units to look for job opportunities in other sectors, leading to the decline of dictionary making. The paper seeks to discuss the future of lexicography and dictionaries in the indigenous African languages of South Africa, particularly in Tshivendḍa, with the focus on challenges and solutions.

Key Words: African language, dictionary, lexicographer, lexicography, missionaries, Tshivendḍa,

Introduction

Tshivendḍa, a minority language in South Africa, is one of the eleven official languages of the Republic of South Africa. Together with the other eight indigenous African languages, it was accorded the status of official language after the democratic elections of 1994; whereas Afrikaans and English had been official languages prior 1994. Lexicography in Tshivendḍa was initiated by the Berlin Lutheran Missionaries who reduced spoken Tshivendḍa into writing. Mafela (2005) maintains that missionaries did a tremendous work in reducing spoken Tshivendḍa to a written language. In many African countries, the missionaries were the first and most active in producing lexicons of African languages (Awak 1990). The Berlin Lutheran Missionaries collected Tshivendḍa terms and compiled term lists containing Tshivendḍa and German terms. The main purpose of compiling term lists was to learn Tshivendḍa. The knowledge of Tshivendḍa helped the missionaries to communicate the gospel to the natives. In this regard, Mandelbaum (1989;37) states:

Thus, to address people in their own native language is to engage in a sacred ritual. You are telling an individual that of all the diverse ethnic ways to express universal feelings, you are electing the way of that individual, of that culture to communicate. Communication then becomes a ritual of reverence for that person's identity and world view. Conversation is possible because the mystery of life is shared in known and familiar ways.

L.T. Marole and N.J. van Warmelo played a major role in compiling the first Tshivendḍa dictionaries in the 1930s. These were followed by lexicographers such as P.J. Wentzel and T.W. Muloiwa. Like in other indigenous African languages, lexicography in Tshivendḍa was not taken seriously by the

authorities at the time. Lexicographers were not trained professionally and did not receive funding from the government. They produced few, but good Tshivenda bilingual and trilingual dictionaries. The dictionaries were meant for learners of Tshivenda and students.

The period after 1994 saw the emergence of lexicography units for all the official languages of South Africa. Each lexicography unit is led by the Editor-in-Chief, assisted by a number of lexicographers. The government, through the Pan South African Language Board, support the units financially. As a result, more Tshivenda dictionaries were produced during this period. Dictionary making was not only focused on bilingual and trilingual dictionaries, but also on monolingual and other types of dictionaries. Awak (1990) says that the types of dictionaries have diversified, some are monolingual, others are bilingual or trilingual, while a few comprise more than three languages. Currently many lexicography units are witnessing the exodus of lexicographers because the financial support from the government does not sustain the units. This is a challenge to the development of indigenous African languages through dictionaries and lexicography. The paper seeks to discuss the future of lexicography and dictionaries in the indigenous African languages of South Africa, particularly in Tshivenda, with the focus on challenges and solutions.

The concepts ‘dictionary’ and ‘lexicography’

It is not easy to draw the distinction between the concepts ‘lexicography’ and ‘dictionary’. When one makes mention of lexicography, one also refers to dictionaries. Lexicography is the process, whereas dictionary is the result of the process. Hartmann’s (1983) distinction between the two concepts is based on their functions. His explanation reads: “The purpose of lexicography is the production of dictionaries, and dictionaries deal among other things with the ever-changing meanings of words; ...” (Hartmann, 1983:3). He therefore defines lexicography as the process of dictionary making which involves the gathering and processing of the word stock, structuring and characterising the material in the corpus, and the preparation of the material that has been collected, sifted and treated for publication (Hartmann 1983).

When he defines dictionaries, he considers what users look up in a dictionary. According to Hartmann (1983), problems which make people turn to dictionaries for help include uncertainties about spelling and pronunciation, curiosity about the origin of a word or expression, the search for suitable synonyms in composition, translating from or into a foreign language, etc. The problems mentioned above are referred to in the definition of a dictionary by Kipfer (1984:1): “A dictionary is a reference book containing the words of a language or language variety, usually alphabetically arranged, with information on their forms, pronunciations, functions, meanings, and idiomatic uses.” In this regard, Landau (1984) mentions that dictionaries often include information about spelling, syllabication, pronunciation, etymology, usage, synonyms, and grammar, and sometimes illustrations as well. In addition to serving as reference books, dictionaries are a record of the vocabulary of a language (Jackson, 2002). The above exposition reveals a linguistic approach towards defining the concept ‘dictionary’.

Nielson (2009) uses the lexicographic approach when defining the concept ‘dictionary’, which focuses on the three significant features of a dictionary that help to shed light on the existence of a dictionary as an object of investigation, description and analysis. Nielson (2009:27) asserts:

Firstly, the overriding feature of a dictionary is that it has been designed to fulfil one or more functions, referred to as *lexicographic functions*, e.g. communicative functions such as the understanding of texts, translation and text production, and cognitive functions such as knowledge acquisition in communication-free contexts. Secondly, the dictionary contains *lexicography data* that has been selected to support the function(s) of the dictionary. ... Thirdly,

the *lexicographic structures* combine and link the data in order to support and fulfil the dictionary function(s).

Both the lexicographic data and the lexicographic structures are there to support the lexicographic functions.

Current status of Tshivenda lexicography and dictionaries

Unlike in the period prior to 1994, post 1994 saw the establishment of national lexicography units for all the official languages, namely, Afrikaans, English, IsiNdebele, IsiXhosa, IsiZulu, Sesotho, Sesotho sa Leboa, Setswana, Siswati, Tshivenda and Xitsonga by the democratically elected government under the guidance of Pan South African Language Board (PanSALB). The main purpose of establishing the units was to help indigenous African languages, which did not receive funding for lexicography work before 1994, to focus on developing and preserving the languages through the production of dictionaries. Muloiwa in the Preface of Tshikota (2006:viii) has this to say:

It is common knowledge that South Africa has elected to use eleven official languages, one of which is Tshivenda. In line with this decision PanSALB deemed it necessary that all the official languages, but more especially those of the Blacks, be developed and expanded in various spheres to empower them for use in different domains of the government, economy, education, etc. One of the means of realising this objective is the compilation of various dictionaries in each of the official languages.

The Tshivenda National Lexicography Unit was established in 2001 under the leadership of the Board of Directors. According to Muloiwa, in Tshikota (2006) the appointment of the staff (Editor-in-Chief and lexicographers) for the compilation of Tshivenda dictionaries marked the commencement of the work in earnest. Since the introduction of national lexicography units, the number of dictionaries and lexicography works increased in the indigenous African languages. Unlike in the period prior 1994, lexicographers in the Tshivenda National Lexicography Unit compiled different types of dictionaries, including monolingual dictionaries, bilingual dictionaries, dictionaries of proverbs, and dictionaries of idioms. The dictionaries are meant for both the learners of Tshivenda and the native speakers. Hereunder are dictionaries and terminology lists produced by Vhavana lexicographers up to date.

Die Verba des Tshivenda (1904) by Theodore and Paul Schellnus

Tshivenda – English Dictionary (1937) by N.J. van Warmelo

Phrase Book (1932) by L.T. Marole

English – Venda Vocabulary (1954) by L.T. Marole

Afrikaans – Venda: Vocabulary and Phrase Book (1955) by L.T. Marole

Phindulano: English – Venda Phrases (1956) by L.T. Marole

Teo dza Tshivenda – Venda Terminologie – Venda Terms (1958) by N.J. van Warmelo

Venda Terminology and Spelling, No. 2 (1962) by the Department of Bantu Education

Venda Terminology and Orthography No. 3 (1972) by the Department of Bantu Education

Verklarende Woordelys: Volume 1 (1972) by Joubert and Rapea

Verklarende Woordelys: Volume 2 (1972) by Joubert and Rapea

Trilingual Elementary Dictionary: VENDA – AFRIKAANS – ENGLISH (1976) by P.J. Wentzel and T.W. Muloiwa

Dictionary of Basic English: Venda (1984) by K.B. Hartshorne

Ifa Lashu la Maambele (1987) by M.C. Neluvhalani

Venda Dictionary: Tshivenda – English (1989) by N.J. van Warmelo

Multilingual Mathematics Dictionary for Grade R to 6 (2003) by the Department of Arts and Culture

Multilingual Mathematics Dictionary for Grade 1 to 6 (2003) by the Department of Arts and Culture

Multilingual Natural Science & Technology for Grade 4 to 6 (2005) by the Department of Arts and Culture

Multilingual Glossary of Health/Medical Terminology (n.d) by the Department of Sport, Arts and Culture – Limpopo Province

Tshivenda – English English - Tshivenda Bilingual and Explanatory Dictionary (2006) by Tshivenda National Lexicography Unit

Tshivenda – English Dictionary of Proverbs (2010) by Tshivenda National Lexicography Unit

Tshivenda – English Dictionary of Idioms (2010) by Tshivenda National Lexicography Unit

Thalusamaipfi ya Luambo Luthihi ya Tshivenda (2010) by Tshivenda National Lexicography Unit

Thalusamaidioma ya Luambo Luthihi ya Tshivenda (2010) by Tshivenda National Lexicography Unit

Thalusamirero ya Luambo Luthihi ya Tshivenda (2012) by Tshivenda National Lexicography Unit

Tshivenda – English English - Tshivenda Dictionary of Proverbs (2012) by Tshivenda National Lexicography Unit

Thalusamaipfi ya Luambo Luthihi ya Tshivenda (2015) by Tshivenda National Lexicography Unit

Tshivenda – English English - Tshivenda Bilingual Dictionary Mathivha-Milubi-Maḍadzhe Edition (2015) by Tshivenda National Lexicography Unit

Many of the lexicography works produced before 1994 are terminology lists, with a few dictionaries published by individual lexicographers. The Tshivenda National Lexicography Unit published nine dictionaries within a short period. The unit managed to achieve this number because it had the capacity. The past few years saw lexicographers leaving lexicography units for other job opportunities because they cannot sustain themselves financially. The government grant is not enough to run the units. As a result, two members of the Tshivenda National Lexicography Unit, including the editor, left for other careers. This could affect the quality and quantity of lexicography works produced. The lack of enough funding is a real challenge to all lexicography units of South Africa. The work is left to a few lexicographers who find it difficult to gather data, sift it and compile dictionaries.

The future of Tshivenda lexicography

The future of Tshivenda lexicography does not look good even if the government is encouraging the development of former marginalised languages. Lexicographers are leaving national lexicography units for a number of reasons, emanating from lack of enough funding. The government, which supports the

lexicography units financially, does not allocate enough funds for carrying out the activities of the units. This is a challenge, especially for the indigenous African languages. Such a challenge leads to a decline of the production of lexicography works. Some lexicographers find it logical to look for other job opportunities in other sectors for the purposes of saving the lexicography units. As indicated above, the Tshivenda National Lexicography Unit lost the editor and a lexicographer as a result. The remaining lexicographers find it difficult to cope with the workload. It is also impossible for them to attend relevant workshops and conferences to update themselves with new developments in the field.

After more than fifteen years of the existence of the Tshivenda National Lexicography Unit, it is expected that it can now sustain itself through the proceeds from dictionaries produced during the years. Instead, it gets very little from these outputs. Perhaps it does not have good marketing strategies. However, the poor culture of dictionary use seems to be the main problem. Vhavana would rather use English and Afrikaans dictionaries instead of Tshivenda dictionaries to find information about languages. They take it for granted that because they are speakers of the language, they do not need a dictionary. This attitude leads to the unit earning very little from their dictionaries because they are not selling well. If they were selling well, the unit would be able to add the proceeds to the government grant to sustain itself. Many people do not even know that there are dictionaries in Tshivenda or that the Tshivenda National Lexicography Unit exists. They cannot search for Tshivenda dictionaries in bookshops because they consider them non-existent. It is surprising to find that non-speakers of Tshivenda, who want to learn the language, are aware of the existence of the Tshivenda dictionaries.

A lexicography unit cannot solely rely on the government grant to sustain itself, it must fundraise. The Tshivenda National Lexicography Unit relies on the government grant and royalties to run its activities. It does not market itself to the citizens to attract some donors. There are people who are passionate about the development and preservation of the language who will be interested in donating funds for the benefit of the language. Dictionaries and lexicography work in general develop and protect a language.

Individual lexicographers have left the work of producing dictionaries to the hands of the Tshivenda National Lexicography Unit because dictionary making is now considered the responsibility of the government. This attitude does not contribute towards the development of the language. Publishing companies also do not help in this regard because they do not encourage the production of dictionaries. They are reluctant to be involved in the production of dictionaries which do not give them a good return. Instead they focus on publishing creative works which are meant for school children and give them a good return. Lexicographers at tertiary institutions focus much on the theory of lexicography on the expense of practical lexicography. They teach lexicography to students without initiating dictionary projects which at the end can produce dictionaries like it used to be done prior 1994. Professor P.J. Wentzel and Professor T.W. Muloiwa of the University of South Africa, for example, worked on a dictionary project and produce a Tshivenda dictionary which is still in demand today.

Producing a dictionary involves a lot of activities which require capacity and enough funding. Staff, free-lance or in-house, is needed to keep the activities of the project going. Dictionary making requires the involvement of people with a range of specialist knowledge and skills (Jackson 2002). Some staff members' responsibility is to gather data for the compilation of the dictionaries. Word list must be selected and sources for definitions must be identified. Experts are consulted to provide definitions in special fields such as medicine and economics. Headwords are defined and entries are proofread and edited. The costs of office space, equipment, administration and employee fringe benefits makes a dictionary project costly (Landau 1984). Producing a dictionary also involves design specifications, papers, printing and binding, i.e. if it is a paper dictionary. All the above mentioned require good financial support. Jackson (2002:161) writes: "Any dictionary, apart perhaps from the occasional

scholarly undertaking, is a commercial venture. It requires considerable investment in staff, equipment, materials and time.” If dictionary making in Tshivenda is to survive the challenges it is facing, solutions must be found.

If the government is really serious about developing and preserving all the official languages, it must increase grants for lexicography units, especially those linked with the indigenous African languages. This opportunity must also be extended to individual lexicographers who are interested in involving themselves in lexicography.

The Tshivenda National Lexicography Unit must not only rely on government grant for its survival, it must fundraise. The Afrikaans and English lexicography units are doing well in this regard and they remain strong throughout. This can be achieved if the Vhenda are proud of their language and ready to support its development financially.

Dictionary making should not be the sole responsibility of the Tshivenda National Lexicography Unit. Other stake holders such as individual lexicographers, publishing companies, and tertiary institutions must involve themselves in dictionary making for the purpose of developing and preserving the language and culture. For example, the University of Venda, which hosts the Tshivenda National Lexicography Unit, must co-support the unit financially with the government. The unit can become a great research asset for the university because it will produce many lexicographic research outputs. Publishing companies must plough back to the communities, profits earned, by supporting the development of indigenous African languages through dictionary making.

Conclusion

The exposition above dealt with the future of dictionary making in Tshivenda. The discussion was introduced by sketching the background information about Tshivenda lexicography. The current status of Tshivenda lexicography reveals that the Tshivenda National Lexicography Unit is the sole dictionary making body post 1994, and that lexicographers are leaving the unit because of financial challenges. It is further revealed that financial challenges are caused by a number of factors, among others, lack of enough funding from the government, lack of good marketing strategies, failing to fundraise and non-participation of stakeholders in dictionary making. The paper concludes by recommending measures that should be taken to make the future of Tshivenda lexicography look better.

References

- Awak, Mario Kidda. 1990. Historical Background, with Special Reference to Western Africa. In Hartmann, R.R.K. *Lexicography in Africa: Progress Reports from the Dictionary Research Centre Workshop at Exeter, 24 – 25 March 1989. Volume 15*. Exeter Linguistic Studies: University of Exeter Press. 8 – 18
- Hartmann, R.R.K. 1983. On the Theory and Practice: Theory and Practice in Dictionary-making. In Hartmann, R.R.K. (ed.). *Lexicography: Principles and Practice*. London: Academic Press. 3 – 11.
- Jackson, Howard. 2002. *Lexicography: An Introduction*. London: Routledge.
- Kipfer, Barbara Ann. 1984. *Workbook on Lexicography: A Course for Dictionary Users with a Glossary of English Lexicographical Terms*. Exeter Linguistic Studies: University of Exeter.
- Landau, Sidney, I. 1984. *Dictionaries: The Act and Craft of Lexicography*. New York: Charles Scribner's Sons.
- Mafela, M.J. 2005. *Tshivenda Literature: A Historical Sketch with Special Reference to its Bibliography*. Pretoria: Khande Publishers.

Mandelbaum, J. 1989. *The Missionary as a Cultural Interpreter*. New York: Peter Lang.

Nielson, Sandro. 2009. Reviewing Printed and Electronic Dictionaries: A Theoretical and Practical Framework. In Nielson, Sandro & Tarp, Sven (eds.). *Lexicography in the 21st Century: In Honour of Henning Bergeholtz*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 22 – 41.

Tshikota, S. (ed.). *Tshivenda – English English - Tshivenda Bilingual and Explanatory Dictionary*. Cape Town: Phumelela Publishers.

TOWARDS DIGITAL DICTIONARIES HAVING MORPHOLOGICAL ANALYSIS

Mutee U Rahman

Isra University

Tafseer Ahmed

Mohammad Ali Jinnah University

Abstract

We present a framework and a tool for augmenting morphological analysis in digital dictionaries. The system enables us to analyze lexical tokens and generate the corresponding morphological structure and features. Instead of listing all the morphological forms (including derivations) and compounds of a lemma in the dictionary, the system analyses the input tokens using finite state transducers and generates the corresponding morphological structure. We also provide a graphical user interface (GUI) to input lemma list, inflections, templates and paradigms to generate/analyze different word forms/compounds. The GUI helps a lexicographer to create lexical transducers without knowing the syntax of finite state systems. Currently, we present the solutions for Urdu, Sindhi and Punjabi, however it can be extended/used for other morphologically rich languages e.g. Turkish and Arabic etc. In contrast to existing frameworks we present a configurable system which enables the user to define inflection rules, templates, derivations, and paradigms. Also, the focus in this implementation is given to internal structure (represented in attribute value matrix style).

Key Words: Digital Dictionaries, Morphology, Lexicography, Morphological Analysis

1. Introduction

Most of the digital or machine-readable dictionaries are based on ordinary paper based dictionaries which inherit problems like omitting the essential lexical and linguistic information, unsyntactic compilation, and ambiguities in lexical entries (Atkins & Levin, 1991). The reason behind this problem is that conventional dictionaries were primarily written and compiled for human use (Boas, H.C., 2009). Intended end user was a person using a computer for dictionary lookup / search. Paper based dictionaries do not list the morphological forms of a word. It is presumed that the user knows about the word formations from root/stem words. Many of the online dictionaries do not have morphological forms, for instance, consider the example of two dictionaries (that can be termed as largest paper based and online dictionaries of Urdu and Sindhi). Urdu Dictionary Board has an online version of its 22 volume dictionary at url <http://www.udb.gov.pk/>. The dictionary does not support morphological forms. If a user searches the word لفظ (lafz meaning ‘word’), its dictionary entry is retrieved. However, if the user searches for its plural لفظوں (lafzon meaning ‘words’) then the system cannot find its dictionary entry. The search system or lexical

entry representation is unable to relate it with the entry لفظ. Same is the case with online dictionary of Sindhi at url <http://dic.sindhila.edu.pk/> compile by Sindhi Language Authority. The dictionary entry of لفظ (lafz meaning ‘word’) is available, but it does not recognize its plural لفظن (lafzan meaning ‘words’). This creates problem for non-native language learners and computer applications which do not know the morphological rules of the language.

With the advancement of technology, dictionaries with rich lexical entries including different features in machine readable eXtensible Markup Language (XML) format are developed (Litkowski, K., 2005). Though, these dictionaries provide lexical information in machine readable format with efficient search facilities yet most of these dictionaries list the base forms of head words (singular stems in case of nouns, present tense form in case of verbs etc.) with their part of speech, pronunciation, origin, and meaning. These digital dictionaries when used in modern NLP systems are unable to analyze the different surface forms of words occurring in a corpus due to the lack of essential lexical / morphological information. For example, an entry for the word “denationalization” is encoded as a noun (singular) or shown as a derived noun from verb denationalize (<https://en.oxforddictionaries.com/definition/denationalization>, <https://www.merriam-webster.com/dictionary/denationalization>). However, the lexical analysis of “denationalization” might have following results.

Word: denationalization

Lexical Analysis:

$$_6[5[de+PP]_5 _4[_3[_2[1[0[nation+NN]_0-\Phi+SG]_1-al+JJ]_2-ize+VB]_3-tion+NN+SG _4 +NN+SG]_6$$

Affixation pattern of word form ‘denationalization’ generated from head noun ‘nation’ is shown here. During this formation various morphemes are combined at different levels forming different word patterns (with same root) which are inflections as well as derivations. For example, at level 2 an adjective ‘national’ is derived from noun ‘nation’ of level 1. Computational lexicons which are used in NLP applications are required to encode essential lexical information as shown in above example entry, this includes stem/head word, POS, inflectional variants, and derivational path along-with lexical and grammatical features. This information becomes more important when we deal with morphologically rich languages. These languages have tens or hundreds of morphological forms corresponding to the words. For example, Turkish verb has large number of morphological forms (see <https://cooljugator.com/tr>) and the Sindhi verb has more than 80 morphological forms (Rahman & Kazi, 2017).

We present a framework and a tool for augmenting morphological analysis in digital dictionaries where user can analyze lexical tokens and generate the corresponding morphological structure and features. Presented framework analyses the input tokens using finite state transducers (Karttunen, 2000) and

generates the corresponding morphological structure without listing all the morphological forms (including derivations) and compounds of a lemma in the dictionary.

There are many tools available for generating the morphological forms of a word and/or morphologically analyzing a word to its root form including PC-Kimmo, XFST, Foma, and OpenFST etc. These tools are very powerful however to use these tools a user or developer need to know about regular expressions and transducers to use these tools. The simplified modules of these tools (e.g. lexc of xfst/foma) are still complex to learn and use for a lexicographer that does not have background in computer programming or scripting (writing a simple programming language).

Hence, we present a framework and a tool for augmenting morphological analysis in digital dictionaries. A graphical user interface (GUI) is provided to define parts of speech (POS) feature space with interface to define POS, their features and feature-values. GUI further enables the user to define POS inflection paradigms along-with their types (Prefix, Suffix, and Template) and features. Inflection paradigms are then mapped to POS definitions. The GUI helps a lexicographer to create lexical transducers without knowing the syntax of finite state systems. System in current state presents solutions for Urdu, Sindhi and Punjabi, however it can be extended/used for other morphologically rich languages e.g. Turkish and Arabic etc. In contrast to existing frameworks like (Attia, et. al, 2011) we present a configurable system which enables the user to define inflection rules, templates, derivations, and paradigms. Also, the focus in this implementation is given to internal structure (represented in attribute value matrix style).

Subsequent sections discuss the details of proposed framework and GUI with examples of morphologically rich south Asian language Sindhi, followed by discussion, conclusion and references sections.

2. Methodology: Augmenting Morphological Information

South Asian languages contain higher number of inflectional and derivational surface/word forms with rich morphological features. Formation of words include number, gender, and case inflections in nominal elements, causative formations in verbs, pronominal suffixation in nouns, verbs, postpositions and adverbs. For example consider a singular, feminine, nominative Sindhi noun form dil “دل” (heart). Conventional dictionaries only list this singular feminine form and do not provide any information about other morphological forms of dil “دل”. Apart from usual number and gender inflections morphological forms nouns in Sindhi are morphologically marked by case and pronominal suffixes. Lexical analysis of a sample form of dil “دل” with ablative case marking and plural number is shown below:

Word: diliyanaan 'دلینان'

Lexical Analysis:

${}_2[{}_1[{}_0[\text{dil+Noun+Fem}]_0\text{-iyan+Pl+Obl}]_1\text{-aan+Abl}]_2$

This is a complex noun entry with different lexical attributes at different levels. Lexical entry is derived from singular feminine noun root, changes its forms at various levels with different lexical attributes and finally a feminine noun with plural number and ablative case is formed. Existing Sindhi dictionaries / digital dictionaries do not list these types of formations. Neither they provide any information about different affixes (-iyan, -aan etc.) and their attributes or how these affixes are used to form words (orthography). It is presumed that end user / native speaker knows about the derivation/inflection and word

formation rules. Word formations are further complicated by nonconcatenative morphology and irregular inflections.

The morphological structure of a word is represented by using a new proposed scheme Morphtrix (Ahmed & Rahman, 2019) where part-of-speech and its morphological features are encoded along-with word's internal structure containing morphemes, affixes and templates in a hierarchical manner. Different attribute value relations and their hierarchies are represented by using attribute value matrix (AVM).

AVM representation of Sindhi noun form dilyanaan "دلینان" is shown in Figure 1. AVM shows surface form "دلینان", its transliteration, POS, and morphological features (MF) which show that this is a feminine, plural, ablative formation. The structure further represents the stem form and its inflectional hierarchy showing that stem of "دلینان" is "دلین" which is feminine, plural, oblique form. This hierarchy of stem formation continues to root a feminine noun dil "دل" (heart). The set entry Affix or Template lists the pronominal suffixes with their attributes at various levels of stem formation providing information about different morphemes being used in stem formation with their feature values. A user interface is designed to input root words their feature values, inflectional and derivational paradigms, and mapping of these paradigms with root words or stems to have surface form words generated by finite state transducers. System uses these transducers either to generate surface forms with necessary AVM structure or to have morphological analysis of surface form words. This is possible due to reversible nature of finite state transducers. Details of GUI designed for entries and mappings are discussed in following section.

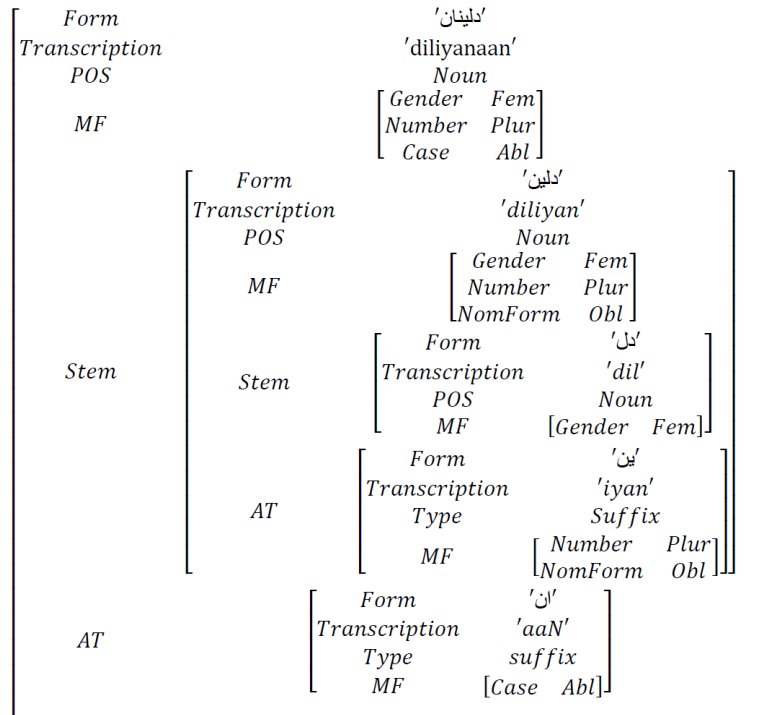


Figure 8. Attribute Value Matrix showing an entry of a plural ablative Sindhi noun.

3. Methodology: Roots, Inflection and Derivation Paradigm Mappings

As discussed above a web based GUI is designed for setup entries of roots, inflections, derivations and their paradigm mappings and to generate surface forms. GUI environment provides definition and customization facility to the user to setup lexical entries environment of his/her choice. Our discussion in following sections will be with different examples of Sindhi Noun morphology as Sindhi is a morphologically rich language with various types of morphological constructions. Nouns in Sindhi are morphologically marked by number, gender, case, and can have pronominal suffixes as well. Following subsections discuss feature space setup, inflectional paradigm definitions, entering roots and their paradigm mappings, and generation of derived word forms by applying derivational morphemes.

3.1. Feature Space Definitions

Figure 2 shows the environment of POS feature space setup where user can enter list of parts of speech, their features, and feature values. The screenshot of Figure 2 shows highlighted entries of Noun its features (Number, Gender, Case, and Form) and feature values (Singular and Plural) of selected feature number. Noun part-of-speech is defined with number, gender, case and form features and their respective feature values singular and plural values of number for instance. Optionally user is allowed to import the list of POS along-with feature and feature values from XML or CSV files.

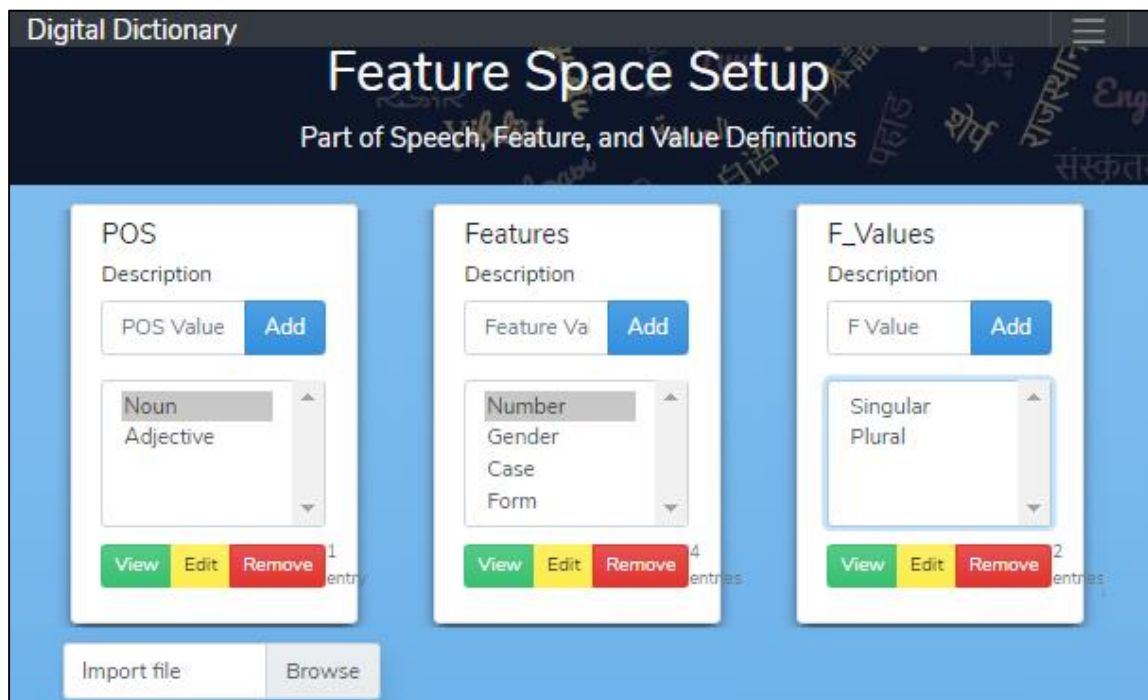


Figure 9. Feature Space Setup Environment.

3.2. Defining Inflection Paradigms and their Mappings

As discussed earlier gender, number, case etc. cause inflectional changes in Sindhi nouns. These inflections are defined by different inflectional paradigms based on type / sub-type of nouns. For example a feminine singular noun will have different inflectional paradigm as compared to masculine singular noun of same type and a feminine singular with non-yE ending noun of one category will have different inflectional paradigm than a non-yE feminine noun of another category. Inflection paradigm setup allows the user to

define different inflectional paradigms, their types along-with features and map these paradigms to respective part-of-speech. Figure 3 shows setup of a Sindhi feminine noun paradigm named N_Cat_1_a with four entries of different suffix type inflections. After feature space setup this is second stage where different paradigms are defined for various POS classes. Next step it to map these paradigms to different entries of root words.

3.3. Root Words to Paradigm Mapping

Feature space and paradigm definitions discussed above need to be mapped to generate surface form word list. This root paradigm mapping interface is shown in Figure 4. User can add single root entry in the roots / lemmas list or optionally upload list of root words and then select the desired root word to define its POS and paradigm mapping. GUI allows the user to add, edit and remove these mappings. Figure 4 shows the mapping of three different noun categories. GUI further enables the user to generate surface form words from these root paradigm mappings. Surface form word list can be exported to a text or XML file.

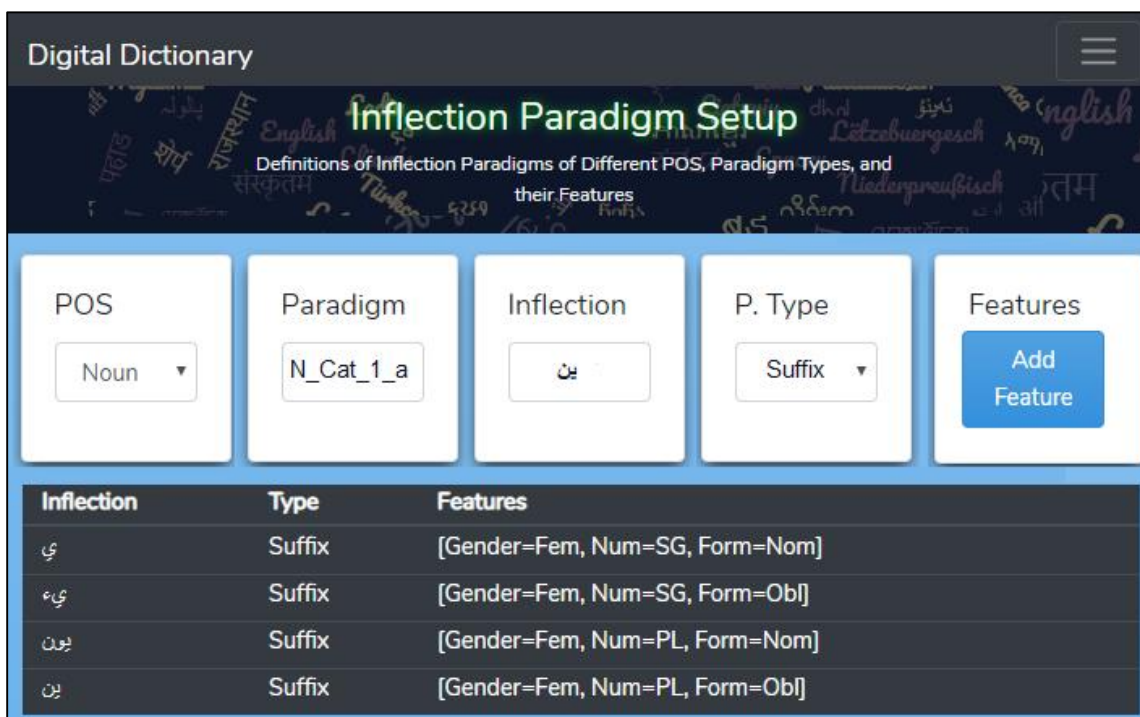


Figure 10. Inflection Paradigm Setup Interface

3.4. Derivational Morphemes and their Mapping with Root Words / Lemmas

Same like inflectional paradigm definitions and their mapping with root words (discussed above) GUI interface allows user to map roots to derivational morphemes. User is enabled to define derivational morphemes, their type (Prefix, Suffix, or Template), source POS, destination POS and features. Interface is almost identical to inflectional paradigm mappings interface and is not shown due to shortage of space. Once user is finished with roots to derivational morpheme mapping (s)he can generate derived word list which can be exported to an XML or text file. Sample mapping entries having derivational morpheme, its type, source POS, target POS and optional list of feature-values are given below:

bAn - suffix - Noun - Adj - []

ee - suffix - Noun - Adj - [Num=SG, Gend=Fem]

aa - suffix - Verb - Veb - [caus='aa']

Some generated derived word form from above mapping of morpheme “bAn” entries include:

Morpheme	Source POS: Lemma	Target POS:	Derived
	Noun : mEz ميز	Adj : mEzbAn ميزبان	
bAn بان	Noun : dar در	Adj: darbAn دربان	
	Noun : nigha نگه	Adj: nighabAn نگهبان	

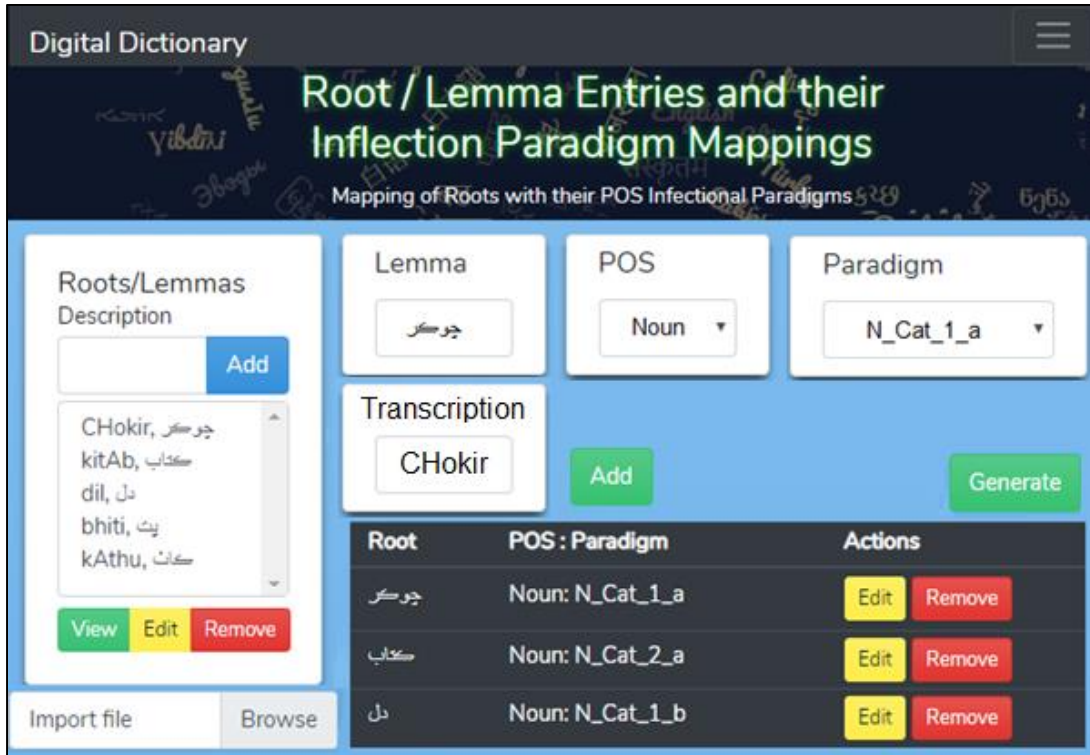


Figure 11. Roots and Inflectional Paradigm Mappings

4. Discussion

The proposed framework and GUI based tool enables the end user which is either a lexicographer or NLP resource developer to define lexicons with rich morphological representations. Dictionaries or morphological analyzers usually dont represent the morphological construction path of a word form (surface form). However, this information is essential to find out the relationship between words and its morphological or derived forms. Finding these relations have applications in natural language processing,

understanding, and generation systems. Also, this information plays a key role in language learning. Existing morphological lexicons (Attia, 2011), (Rahman & Kazi, 2017) are special purpose language specific lexical resource developments. Apart from developed linguistic resources there are tools available to develop such resources base on finite state technology. These tools include PC-Kimmo, XFST, Form, and OpenFST. However, despite of the fact that these tools are very powerful and computationally tractable; one need to be a computer programmer / programming script writer to work on these tools. Linguists / lexicographers and NLP resource developers do not feel comfortable to work on these tools easily. Our proposed system tries to solve this difficulty of knowing tedious programming details by providing a GUI based system to define customizable lexicons with rich inflectional and derivational morphological constructions. The proposed system not only provides a user friendly GUI to define morphologically rich lexicons but also represents the details of word formation path from surface form word to roots and vice versa. To represent the complex morphological entries an AVM based model (Ahmed & Rahman, 2019) is used. This AVM structure represents the word formation hierarchies along-with morphemes and attributes at different levels of word formation. A sample attribute value matrix can be seen in Figure 1. System further enables the user to define parts-of-speech, their features, and feature values. This is called feature space definitions. In next step user is provide a GUI to define inflectional paradigms of various POS classes defined in first step. Subsequently user need to define the lexical entries their part-of-speech and inflectional paradigm mappings. For all these steps user don't need to be an expert in programming. If the user just has basic concepts of linguistics and word morphologies then (s)he will be able to define all these setup configurations. Once POS feature space, paradigms, and lexion entries and their mappings with paradigms are configured surface form words can be generated just by single click. This process is discussed in sections 3.1 to 3.4 and shown in figures 2 - 4. The system in current state provides the basic functionality but still in development phase, future developments will provide more robust GUI with generic configurations to be used for NLP resource development, lexicography and linguistic research.

5. Conclusion

Augmenting morphological analysis in digital dictionaries is not only important for NLP applications but has applications in lexicography and language learning as well. Existing digital dictionaries do not provide sufficient lexical and morphological information to end users (humans or NLP based software applications). Proposed framework and GUI based application is a platform to develop morphologically rich lexical resources to be used by linguists, lexicographers, and natural language processing resource developers without knowing the finite state technology and programming details. While developing the framework and GUI focus is given to south Asian languages Sindhi, Urdu-Hindi, and Punjabi due to their rich inflectional, derivational, and template based morphology. System provides facility to define customized feature spaces, and paradigms and is extendable to other languages.

References

- Ahmed, T. and Rahman, M. (2019). Morphtrix: A Comprehensive Scheme to Encode Morphological Structures. Submitted to 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology co-located with ACL 2019 Florence, Italy, to be held on August 02, 2019.
- Atkins, B.T.S. (1993) The contribution of lexicography. In: Bates, M. and R.M. Weischedel (eds.), *Challenges in Natural Language Processing*, 37–75. Cambridge: Cambridge University Press.
- Atkins, B.T.S. and B. Levin 1991 Admitting impediments. In: U. Zernik, (ed.), *Lexical Acquisition Using Online Resources to Build a Lexicon*, 233–262. Hillsdale: Lawrence Erlbaum Associates.

- Attia, M., Pecina, P., Toral, A., Tounsi, L., & van Genabith, J. (2011). An open-source finite state morphological transducer for modern standard Arabic. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (pp. 125-133). Association for Computational Linguistics.
- Boas, H. C. (Ed.). (2009). *Multilingual FrameNets in computational lexicography: Methods and applications* (Vol. 200). Walter de Gruyter.
- Butt, M., & King, T. H. (2007). Urdu in a parallel grammar development environment. *Language Resources and Evaluation*, 41(2), 191-207.
- Dictionary, M. W. (1996). Merriam-Webster. Incorporated, 10th edition edition.
- Humayoun, M., & Ranta, A. (2010). Developing punjabi morphology, corpus and lexicon. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*.
- Karttunen, L. (2000). Applications of finite-state transducers in natural language processing. In *International Conference on Implementation and Application of Automata*(pp. 34-46). Springer, Berlin, Heidelberg.
- Litkowski, K. (2005). Computational lexicons and dictionaries. *Encyclopedia of Language and Linguistics*, 2, 753-61.
- Rahman, M., & Kazi, H. (2017). Developing a Computational Syntax of Sindhi Language in Lexical Functional Grammar Framework. *Sindh University Research Journal-SURJ (Science Series)*, 49(4), 733-738.

CONNOTATION VERSUS DENOTATION: THE EFFECTS OF CONNOTATION IN THE LEMMATIZATION OF TERMINOLOGY

Dr MV Mojela

University of Limpopo

Polokwane, Rep of South Africa

Abstract

The research analyses the distortions of the denotative meaning of words such as ‘Bantu’ and ‘African’, as used in the South African context as a result of the historical connotations which are associated with the words in South Africa. The philological and the lexicographic definitions of the words create major challenges to both the South African lexicographers and the philologists because of the disparities it creates when compared with the meaning of the words as defined globally. Dictionary.com (2018) defines the term, connotation, as follows:

Connotation refers to a meaning that is implied by a word apart from the thing which it describes explicitly. Denotation or denotative meaning is defined by Dictionary.com (2018) as: *The explicit or direct meaning or set of meanings of a word or expression as distinguished from the ideas or meanings associated with it or suggested by it.* According to Dictionary.com (2018), the Bantu languages are part of the Southern Bantoid branch of the Benue–Congo, which is one of the language families grouped within the Niger–Congo phylum. The word was first coined by Malcolm Guthrie in his ‘Proto Bantu hypothesis’ which tried to determine the origin of this sub-family of languages within the African continent. Despite this classificatory reference of the word, i.e. its denotative reference, the word, Bantu, came to be associated with negative historical connotations which resulted in it being disliked by many people in South Africa, especially the indigenous Black communities. The word came to be associated with colonialism and racism, especially as a result of the word being used in the classification of the South African population into two groups. i.e. Europeans, referring to the white people and the Bantu, referring to the black indigenous communities. As a result of these negative connotations, the word is often substituted by the word ‘African’, especially when referring to the languages, as a way of avoiding to call the indigenous languages ‘Bantu languages’, which is also a further complication of ambiguity in the classification of the languages of Africa. By African languages the lexicographers and philologists refer to all the languages spoken within the African continent, which the Bantu is part of the language families. The lemmatization of these words, Bantu and African, create major confusions because the words are not only defined and lemmatized in South Africa by the South African lexicographers only, but are also lemmatized worldwide by the international lexicographers in all the continents of the planet Earth.

Key Words: African Languages, Bantu Languages, Connotation, Denotation, Lemmatization

INTRODUCTION

This research is aimed at determining the extent to which the connotative meaning of the words corrupts the original or the denotative meaning of the lexical items in a language, as a result of the association of the word with local socio-political events. As a results, these connotations lead to the lexicographic disparities where the words are not given the same definitions by the various lexicographers worldwide. These lexicographic disparities lead to confusions for both the lexicographers and the philologists in the classifications and the lemmatization of the words. Using the word, Bantu, as example, the word is internationally and globally defined with reference to its denotative meaning which pertains to its classificatory reference within the language families in the African continent. Dictionary.com (2018) gives the denotative definition of the Bantu languages as the languages which form part of the Southern Bantoid branch of the Benue–Congo, which is one of the language families grouped within the Niger–Congo phylum. This denotative reference of the word Bantu has metaphorically shifted to acquire negative connotative meanings which are associated with racism due to it being used to refer to the Black people in the racial classifications of the South African population by the Apartheid system. While the secondary objective of this research is basically to disambiguate the meaning of the word ‘Bantu’ and that of the word ‘African’, the primary objective of this research is also to highlight the disparities which are often experienced by the lexicographers with regard to the lemmatization and the definition of the lexical items when defined from one country or one area to another. It is obvious that the lexicographic definition of the word ‘Bantu’ outside the borders of the Republic of South Africa will not have exactly the same meaning as those of the South African lexicographers because of the unique connotations which are attached to the word within South Africa.

METHOD

This research is conducted implementing a combination of relevant methods such as comparative methods, especially the methods which are in line with the philological comparison used by scholars like Malcolm Guthrie (1948) and Joseph Greenberg (1962), in their classification of the languages of Africa to distinguish the Bantu Languages from the other African Language families. In order to get the general background knowledge of the South Africans regarding the classification of the languages of Africa and how the Bantu language sub-family is located within these languages, few oral interviews were held among the students and the academics in the University of Limpopo, Polokwane. This interview method helped to determine the extent to which these groups, i.e. the students and the academics, identify the words ‘Bantu languages’ from the African languages, especially with regard to their denotative meanings. Questions such as the following were asked from the students and academics and their answers captured orally on a cellular phone.:

- What is the meaning of the word ‘Bantu’?
- Who are the Bantu people?
- What is a Bantu language?
- Where are the Bantu languages spoken within Africa?
- What is the difference between the Bantu Language and the African Language?

RESULT

The results of the investigation revealed that lack of proper understanding of the meanings of the words, 'Bantu' and 'African' resulted from the corruption of the denotative meaning of these two words, which resulted from the association of the word 'Bantu' with the negative connotations which did not form part of its original meaning. The interviews conducted within the University of Limpopo gave clear indications that almost 90% of both the student and the academic communities did not know the original or the denotative meaning of the word 'Bantu', which pertains to the classification of the languages of Africa, i.e. a sub-family of the Southern Bantoid branch of the Benue–Congo group. The research revealed that most of the students and the academia believe the word to be one of those racist terms or names coined by the European colonists as a way of undermining the African indigenous communities. It is only the few linguistic scholars and the academia, who had extensive knowledge of the classification of the languages of Africa, who understood the real meaning of the word 'Bantu' and who did not have any problem with the use of the word in a language. The most interesting fact about the word 'Bantu' is that the word is recognized and acceptable to all the Bantu communities outside South Africa. This is seen when most of the international scholars from outside South Africa successfully publish scientific research publications and present academic research papers on the Bantu languages in the international conferences which are held both outside and inside South Africa. These international scholars use the term Bantu freely without any reservation and without any fear because they did not know or experience the negative connotations associated with the word in South Africa.

DISCUSSION

The primary objective with this research is also to highlight the various lexicographical and philological challenges emanating from the attachment of connotations to words which subsequently dominate and prevail over the original or the denotative meaning of the words in a language under the influence of the local or regional activities. In lexicography, both the denotative and the connotative meanings of the words are taken into consideration when defining the lemmas in order to give the dictionary users the comprehensive meaning of the words. To define words like 'Bantu', the lexicographer need to have additional background knowledge of the connotative meanings of the word so that these meanings can all be included and identified when defining the word in the dictionary. This system of defining, or lemmatizing words like 'Bantu' gives additional challenges to the lexicographers, like the following:

- (i) It will not always be possible for the lexicographers to know all the additional connotative references of each and every word in a language
- (ii) The connotative meanings of the words will always differ from one area to another, or from one country to another as against its denotative meaning.
- (iii) The word which is derogative, taboo or unwanted in one area or one country will be acceptable in another country.
- (iv) Connotations increase or are added to words from time to time due to various factors and these changes need the attention of the lexicographer to keep on updating the meanings of the lexical items regularly in the dictionaries.
- (v) Frequently used connotations sometimes tend to dominate, supersede or prevail over the denotative meaning of the words and ultimately replace the original meanings of words.

With reference to the first challenge above, i.e. ‘(i) It is not always possible for the lexicographers to know all the additional connotative meaning of each and every word in a language’, it is discovered that various lexicographers often agree, or give a more or less the same definition, with regard to the denotative reference of the words while they often differ with regard to the connotative reference of the words. These differences seem to depend much on the frequency of usage and the influence of the environment in which the word is lemmatized. For instance, the American Heritage Dictionary of the English Language (2019) defines the word ‘Bantu’ as follows:

1. A member of any of a large number of linguistically related peoples of central and southern Africa.

2. A group of over 400 closely related languages spoken in central, east-central, and southern Africa, belonging to the eastern branch of the Benue-Congo group of the Niger-Congo language family and including Swahili, Kinyarwanda, Kirundi, Zulu, and Xhosa

Collins English Dictionary (2019), gives the following denotative definition for the word ‘Bantu’, but with additional connotative description:

1. (Languages) a group of languages of Africa, including most of the principal languages spoken from the equator to the Cape of Good Hope, but excluding the Khoisan family: now generally regarded as part of the Benue-Congo branch of the Niger-Congo family

2. (Peoples) taboo. South African; a Black speaker of a Bantu language.

The two dictionaries agree on the basic or the denotative reference of the word ‘Bantu’, i.e. the classification of the language as member of the Benue-Congo branch of the Niger Congo family. But Collins English Dictionary has added the fact that the word is taboo in South Africa, even though the dictionary did not indicate the connotative reasons which lead to its unacceptability to the South Africans. It is obvious that the compilers of the American Heritage Dictionary did not know any other additional connotation attached to the word as it is the case in South Africa, while the lexicographers of the Collins English Dictionary knew the negative connotations attached to the word in the South African context. The few examples of the negative connotations associated with the word, Bantu, in South Africa, include the following senses which are associated with racism and the ‘Apartheid’ system, i.e. the word is regarded to mean:

- Black people, as against white people
- People of lower status
- People who worked for the whites
- People who were not allowed to vote, and who were ruled by the white minority
- People who lived in the so-called Bantustans
- People who were given inferior education called Bantu Education
- People who endured severe socio-political oppression, etc.

The following is a map showing the classification of the languages of Africa with the Bantu as one of the sub-families of the Niger-Congo Language Family. It is important to realize that the Bantu population in South Africa form just a tiny minority of the entire group of the Bantu communities who are found outside the borders of the Republic of South Africa



Map showing the classification of the language families of Africa with the Bantu as sub group of the Niger-Congo family. (Wikipedia. 2019, The free Encyclopedia)

All these references which gave the word, Bantu, its additional connotative meaning were not known or felt outside the borders of South Africa and as a result, the word is freely and innocently used outside South Africa just like it was originally coined and used without any stigma by Malcolm Guthrie in his publication, ‘Comparative Bantu’. These arguments confirm and validate items (ii) and (iii) mentioned above, i.e. ‘(ii) the connotative meanings of the words always differ from one area to another, or from one country to another as against its denotative meaning’ and ‘(iii) the word which is derogative, taboo or unwanted in one area or one country is acceptable and freely used in another country’

With regard to the last two items or challenges, i.e. ‘(iv) Connotation increases from time to time due to various factors’, and ‘(v) The connotative meaning sometimes supersede, dominate or prevail over the denotative meaning of the word, which sometimes lead to the loss or changes in the original meaning of the word’ the definitions given by these two dictionaries show that the original meaning of the word ‘Bantu’ only referred to its classificatory role, as in Guthrie’s hypothesis, which is its denotative meaning. But with the course of time the word came to acquire additional connotations which ultimately led to the extension of its meaning in South Africa. This gradual disappearance or loss of the denotative meaning of the word

is usually due to the fact that the majority of the people of South Africa, who are not literate enough about the classification of the languages of Africa, know the word to refer to the non-white population of South Africa. The word as it is used today in South Africa refers to all the non-white communities, including the Khoisan communities, who were never classified as Bantu by scholars like Guthrie, Greenberg, etc.

The second most important objective with this research is the critical analysis of the challenges confronting lexicography with regard to the replacement of the word 'Bantu' by the word 'African' in South Africa. The socio-political pressures from the majority of the South Africans compelled many institutions to replace the word 'Bantu' with the word, 'African' in all private and public use as well as in all official documentations, as in the following instances:

- The word 'Bantu' was replaced by the word 'African'
- Bantu language was replaced by African language
- The Department of Bantu Languages in the Universities and all the institutions of higher learning were replaced by the Department of African Languages

The Shorter Oxford English Dictionary (2007) defines the word 'African' as follows:

African: 'native or inhabitant of the continent of Africa (OE)

Pertaining to Africa; belonging to or characteristic of African people

The Dictionary of South African English on historical principles (1996), defines the word 'African' as:

Any person born or living in Africa

A black person of African descent

Any person born or living in Africa

These definitions denote that all the people living in Africa are Africans, and not only those who live in South Africa. Their languages are, therefore, African languages. It is for this reason that the substitution of the word 'Bantu' by 'African' is very much inappropriate and confusing for the lexicographers and the philological scholars. For the philologists and the lexicographers, the word African languages refer to all the languages which belong to the language families of the entire continent of Africa, such as the Nylotic Languages, the Sudanic languages as well as the Benue-Congo-group of languages which is a section of the Niger-Congo family of languages which the Bantu languages form part of. The lexicographers will need to have a full knowledge of the classifications of the African languages to be able to differentiate between these two words, i.e. 'Bantu' and 'African'. The words 'African' languages cannot be synonyms to 'Bantu' languages because the word African languages also include the Khoisan and the Afrikaans languages. Both the Khoisan and Afrikaans do not form part of the Bantu languages in the classification of the African languages.

The most important significance of this research is the fact that it gave the advantages and the disadvantages of the connotative meanings the lexical items in a language. These advantages and disadvantages resulting from the addition of connotative reference come as follow:

- The advantages usually pertain to the increase in the vocabulary of the language, which mostly result from the semantic extensions of words which lead to the creation of new lexical items via polysemy.
- The advantages also pertain to the instances where nice and positive connotations are added to the words via semantic extensions.
- The disadvantage of meaning extensions of the lexical item come as a result of the addition of the negative connotations, i.e. like the meaning extension which occurred to the word 'Bantu' in the context of the way it is used in South Africa.

CONCLUSION

In conclusion, it is important to give a summary of the objective and the significance of this research in the development of lexicography for the South African languages. The following objectives are outlined and analyzed in this research:

- The comparative analysis of the denotative and connotative references in the development of lexicography in the languages with special reference to the meaning of the word, Bantu, in the South African context.
- The effect of connotations in the development of lexicography for the South African languages, with special reference to the meanings of the words 'Bantu languages' and 'African languages' as used in South Africa.
- The disambiguation of the words 'Bantu languages' and 'African languages' with special reference to the South African socio-political context.

The research gave intensive analysis to show how the denotative meaning of the lexical items can be corrupted by the addition of negative connotations in the meaning of the word in a language. The politicization of the word 'Bantu' by the addition of negative connotations led to the deviation of its denotative reference of language classifications to give emphasis to socio-political connotations. The lexicographic effects of connotations to the words is due to the fact that the connotations are too regional or local while the denotative meanings are much global and international. As a result, the connotative meaning of words are mostly defined by the local lexicographers while the lexicographers who are far away might not know the local connotations of the word. The research succeeded in disambiguating the meaning of the word 'Bantu' in order to expose the real meaning of the word, which was created for the sake of the classification of the languages of Africa and not the political objective which is emphasized in the South African context.

REFERENCE

- DSAE (1996) Dictionary of South African English on historical principles. Oxford University Press, London
- American Heritage Dictionary of the English Language (2019) New York
- Collins English Dictionary (2019) Fifth Edition. Houghton Mifflin Harcourt Publishing Company.
- Greenberg, J.H. (1962), The languages of Africa, Stanford, USA

Guthrie, M (1967). *Comparative Bantu*, Greg Press, London

Leech, P. 1978. *Semantics*, Penguin Books, New York.

Mojela V.M. (2007) *Polysemy and Homonymy, A case study of the challenges relating to lexical entries in the Sesotho sa Leboa/English Bilingual Dictionary*, published in *Lexikos* 17 (2007)

Mojela V.M. 1999. *Prestige terminology and its consequences in the development of Northern Sotho vocabulary*. Unpublished doctoral thesis, Unisa, Pretoria

Mojela V.M. 2005. *Standardization and the development of orthography in Sesotho sa Leboa – A historical overview*. In *the standardization of African Languages in South Africa*, by Vic Webb, University of Pretoria, Pretoria

Shorter Oxford English Dictionary (2007), *Oxford University Press* Wikipedia (2018), *The Free Encyclopedia*

BILINGUAL DICTIONARY FOR ACADEMIC WRITING: HEDGES, BOOSTERS AND ATTITUDE MARKERS AS INTERACTIONAL RESOURCES

Neslihan ONDER-OZDEMIR

Bursa Uludag University, School of Foreign Languages

Abstract

Academic writing for publication tends to be challenging not only in a foreign language but also in the native language of the scholars because “academic English is no one’s first language” (Hyland, 2016a, p. 61). Most of the scholars around the world are often under pressure to publish in English and also in their mother tongue. Following Wu (2016) and Gurlek (2016), this preliminary study sets out to build a bilingual specialized dictionary for the three interactional resources as metadiscourse in academic writing as follows: hedges (e.g., perhaps=*belki*), boosters (e.g., clearly=*net biçimde*) and attitude markers (e.g., unfortunately=*maalesef*). Self-compiled specialized corpora were built from discussion sentences of the doctoral theses (n=90) and research articles (n=90) in three disciplines: engineering, medicine and applied linguistics given that discussion sentences and sections are reported to be more challenging compared to other sections in a text, such as introduction, methods and results. In this study, the methodology employed for reliable results was detailed and also sample words obtained from corpora with their translations accompanied by a sentence were provided for the bilingual dictionary.

Key Words: Academic writing, bilingual dictionary, specialized corpora, metadiscourse

Introduction

“*Effective academic writing always carries the [author]’s point of view.*” Ken Hyland (1996)

English is considered lingua franca to keep up with scientific developments and for scholarly publishing to disseminate research in academia (Hyland, 2016a). Thus, many researchers, including native speakers of English and non-native speakers, are working outside English-speaking countries, aim to publish in English for various reasons, e.g. professional development (Onder-Ozdemir, 2014), academic career (Salager-Meyer, 2014), promotion, research grants (Flowerdew, 2000) or because of pressure in academia (Hyland, 2016b).

While publishing, researchers encounter various problems, such as non-standard English, social conditions (e.g., lack of funding), lack of academic resources and culture, when it comes to publishing in English. There are different perspectives on publication problems. Although most of the studies are conducted to find out problems researchers, who are non-native speakers of English, face while writing in English to publish, I should note that the problems are not present only for non-native speakers but also for native speakers given that “academic English is no one’s first language” (Hyland, 2016a, p. 61). Indeed, both native speakers of English and non-native speakers may encounter similar problems while publishing a research article. As Swales (2004, p. 52) noted: “(certain mechanics, such as article usage aside) *au fond* pretty similar to [the problems] typically by native speakers.” To Swales, being a native or non-native speaker of English is not important, but the experience of researchers in publishing, namely being more or and less experienced, is important.

In this study, the term metadiscourse was defined as “writer’s awareness of the reader and his or her need for elaboration, classification, guidance and interaction” (Hyland, 2008, p. 17) in academic writing. In this study, following Vande Kopple (2002), I am in favour of drawing a distinction between propositional meaning (i.e. information about external reality) and metadiscourse (i.e. writer’s awareness of the reader and his or her need for elaboration, classification, guidance and interaction” Hyland, 2008, p. 17). I should highlight that with the bilingual dictionary in this study, which is in the pipeline, we will be able to compare what propositional meanings in the traditional dictionaries and metadiscourse meanings. This dictionary aims to raise awareness about how we can “connect, organize, interpret, evaluate and develop attitude towards” a written text (Vande Kopple, 2002, p. 93).

Instead of using the term “corpus-based”, Wu (2016) used the term “corpus-assisted” bilingual lexicography with quality control criteria with a focus on five types of meanings as follows: conceptual meaning, connotative meaning, cultural meaning, structural meaning and pragmatic meaning. Hence, accurate translation at word/phrase (lexical) level is significant for the quality in lexicography. With a critical eye, Gurlek (2016) discussed strong (e.g. developed based on available bilingual dictionaries) and weak sides of Turkish lexicography (e.g. lack of learner dictionaries/pedagogical dictionaries, for example, a qualified learner’s dictionary of Turkish has not been written yet).

Drawing on the discussions by Wu (2016) and Gurlek (2016), this preliminary study sets out to build a bilingual specialized dictionary to facilitate academic publishing process through focusing on three interactional resources as metadiscourse (Hyland, 2008, pp. 52-53): (i) hedges, (ii) boosters and (iii) attitude markers, which have been detailed below.

(i) Hedges (such as *possible, perhaps, may, might, appear/seem, likely and suggest*) are language devices that highlight the subjectivity in writing to provide opinion rather than a fact or knowledge and open to negotiation. Thus, hedges are a statement based on the writer’s plausible reasoning. In this study, my priority was to highlight the significance of hedging in academic writing.

(ii) Boosters are words (such as *clearly, demonstrate, in fact, it is clear that*) that express certainty in what we say in writing through highlighting the force of propositions.

(iii) Attitude markers are words (such as *agree, hopefully, difficult, interesting and remarkable*) that indicate the writer’s affective attitude to propositions to convey surprise, agreement, importance, obligation and frustration.

Methodology

In this study, the term *specialized corpus* is defined as a collection of texts belong to text-type of genre to examine specialized language (Gavioli, 2005, p. 7) for the bilingual dictionary. The underlying reason why this study focused on the specialized corpus is that specialized corpus is carefully targeted; specialized structures may occur with regular patterning, distribution and also the pedagogical aims about how specialized corpus is used and applied are likely to be easier to define and delimit (O’ Keeffe et al., 2007, p. 198). In this study, self-compiled specialized corpora were built from discussion sentences of the doctoral theses (n=90) and research articles (n=90) written in 2016-2019 in three disciplines: engineering, medicine and applied linguistics and in light of the literature, four groups of criteria were used to design specialized corpora to obtain reliable results (see Aston & Burnard, 1998; Biber, 1993; Flowerdew, 2004; Huston, 2002; Onder, 2011):

(i) The academic texts selected for the content of the dictionary: Bhatia (1993) stated that we should define the genre/sub-genre properly, so the genre could be distinguishable from other genres and also similar in some ways. Drawing on Bhatia (1993), to select the right kind and size of texts, self-compiled specialized corpora were built from discussion sentences of the doctoral theses (n=90) and research

articles (n=90) in three disciplines: engineering, medicine and applied linguistics given that discussion sections are reported to be more challenging compared to other sections in a text.

(ii) The size of the corpora: Because this study was preliminary, small corpora were built from discussion sentences of the doctoral theses (n=90) and research articles (n=90) in three disciplines: engineering, medicine and applied linguistics.

(iii) Representativeness: The academic texts that were included in the corpora aimed to reveal the full range of availability regarding hedging. Thus, journals were chosen considering the impact factor score. For example, for medicine, well-known journal research articles were chosen from the British Medical Journal (BMJ) and also Lancet Global Health; for engineering journal titled Electrical and Electronics Engineers (IEEE)/ACM Transactions on Audio, Speech, and Language Processing and also IEEE Transactions on Neural Networks and Learning Systems; for linguistics English for Academic Purposes and Journal of Second Language Writing were chosen.

(iv) Permanence: To make the bilingual dictionary representative and permanent, the content of this bilingual dictionary will be updated regularly (Huston, 2002).

Results and Discussion

Wu (2016), who argued that corpus-assisted bilingual lexicography needs criteria for the quality of bilingual dictionaries, five types of meaning (i.e. conceptual meaning, connotative meaning, cultural meaning, structural meaning and pragmatic meaning) at lexical level and sentence level were considered as evaluating criteria. Testing practices will be carried out to further prove the effectiveness of Wu's (2016) theoretical proposal, particularly on translation. I should note that because, in this study, bilingual dictionary is built for academic purposes, I do not expect to find out five types of meaning Wu (2016) highlighted.

Two sample sentences from an original research article in medicine in the British Medical Journal (2019) were provided below with two hedges as metadiscourse, i.e. *probably* and *could*, were italicized and underlined below.

Our lack of statistically significant associations between glucosamine use and subtypes of stroke is *probably* because of small numbers of participants in the subtype groups. Several potential mechanisms *could* explain the observed protective relation between glucosamine use and CVD diseases.

In line with Ken Hyland's (1996) findings hedging is most commonly expressed by lexical verbs, epistemic adverbs, epistemic adjectives and modal verbs, which suggests that when the data analysis is over, hedging will outnumber other types of metadiscourse.

In the bilingual dictionary, each dictionary entry, such as *probably*, will be defined and translated and the translations will be checked by the content experts and also translators for the validity of the bilingual dictionary. In the sample sentence above, "*probably*" was hedging.

Definition in English:

probably (adv): used when saying that you think something is true or will happen, although it is not completely definite

Definition in Turkish:

probably (zarf): muhtemelen, belki

A sample sentence from an original research article in engineering in the journal titled *Electrical and Electronics Engineers (IEEE)/ACM Transactions on Audio, Speech, and Language Processing* (2019): “may”

[...] the continuous exposure to steady tones at moderately high pressure levels **may** introduce fatigue.

A sample sentence from an original research article in applied linguistics in the journal titled *Journal of Second Language Writing* (2018): “*it is likely that*”

[...] *To what extent the students heeded these more informal, verbal explanations [were] beyond the scope of this study, but **it is likely that** the official curricular documents and their emphasis on rigid genre boundaries had a strong impact on student understanding because of their permanence and their direct relationship to students’ grades.*

Conclusion

This study is in the pipeline but so far the findings showed that although the corpora consisted of different disciplines for the bilingual academic dictionary, the disciplines seems to be very similar in terms of academic conventions particularly for the use of words/vocabulary while the findings are discussed. Hence, the bilingual academic dictionary may raise awareness among young researchers how to produce an academic text both in English and in their native language.

References

- Flowerdew, J. (2000). Discourse community, legitimate peripheral participation and non-native English-speaking scholar, *TESOL Quarterly*, 34, 127-150.
- Gavioli, L. (2005). *Exploring corpora for ESP Learning*. Amsterdam: John Benjamin.
- Gurlek, M. (2016, June). Turkish Lexicography: From Ottoman lexicography to Turkish dictionaries. *Proceedings of the 10th International Conference of the Asian Association for Lexicography (AsiaLex2016), Advancing Language Teaching through Lexicography and Corpus-building*. Edited by Shirley N. DITA and Wilkinson Daniel Wong GONZALES, pp. 35-43.
- Hyland, K. (2016a). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, 31, 58-69.
- Hyland, K. (2016b). *Academic publishing: Issues and Challenges in the Construction of Knowledge-oxford Applied Linguistics*. Oxford University Press.
- Hyland, K. (1996). Nurturing hedges in the ESP curriculum. *System*, 24 (4), 477-490.
- Swales, J. M. (2004). *Research Genres*. Cambridge: CUP.
- Ozdemir-Onder N. (2014). The role of English as a lingua franca in academia: The case of postgraduate students in an Anglophone-centre context. *Procedia-Social and Behavioral Sciences*, 141, 74-78.
- Vande Kopple, W. (2002). Metadiscourse, discourse, and issues in composition and rhetoric. In E. Barton and G. Stygall (Eds.), *Discourse Studies in Composition* (Vol. 91-113). Cresskill, NJ: Hampton Press.
- Wu, J. (2016, June). Corpus-assisted Bilingual Lexicography and its Evaluating Criteria. *Proceedings of the 10th International Conference of the Asian Association for Lexicography (AsiaLex2016), Advancing Language Teaching through Lexicography and Corpus-building*. Edited by Shirley N. DITA and Wilkinson Daniel Wong GONZALES, pp. 61-69.

TRANSFORMING GLOSSARIES INTO KNOWLEDGE RESOURCES: FRAME-BASED TERMINOLOGY APPLIED TO MILITARY SCIENCE

Pamela Faber

Pilar León-Araúz

Faculty of Translation and Interpreting

University of Granada

Acknowledgements

This research was carried out as part of the project FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. The authors would like to thank the master students who participated in the evaluation of EcoLexiCAT.

Abstract

This paper describes a Frame-Based Terminology approach to the military terminology of the Spanish Armed Forces. The alphabetically organized (PD0-000) glossary of military terms of the Spanish Armed Forces was transformed into *MiliMarco*, a bilingual terminological knowledge base in which each concept appears within the context of a frame that highlights semantic relations and conceptual structure. The objectives of *MiliMarco* included the design and implementation of a frame-based military knowledge resource, information extraction techniques from a corpus of military documents, and the potential contribution to the interoperability of the resource with those of international organisms. Frame-based resources enhance access to domain knowledge in a contextualized way, since embedding concepts in a knowledge structure activates associative information in semantic memory and promotes context availability. The design and population of *MiliMarco* involved the analysis and transformation of the content in the glossary as well as the extraction of new information. For instance, specialized knowledge structures were extracted from the implicit structure of the definitions in the glossary and from the explicit lexicalization of semantic relations in the corpus. New concepts were added based on the gaps encountered in the glossary and new data categories were included, such as images, collocations and contexts. Although still and on-going project, the resulting knowledge base is currently a concept-oriented resource where users can browse through different semantic networks and frames based on their cognitive and communicative needs.

Key Words: Terminology knowledge base, military science, frame-based terminology

1. Introduction

This paper describes a Frame-Based Terminology (FBT) approach (Faber 2012, 2015) to the military terminology of the Spanish Armed Forces. Promoting successful communication in multilingual scenarios evidently entails more than facilitating a standardized list of alphabetically-arranged concepts. Since misinterpreted messages can have dramatic consequences in military settings, it is also necessary for text senders and receivers to have access to the same context. In other words, they must possess the same domain knowledge to facilitate mutual understanding. This objective is easier to achieve when terminological resources are context-oriented or frame-based.

Knowledge of terminological units and their meanings also means being aware of how these units combine with others and in which scenarios these combinations may occur. Users must be able to understand the range of contexts activated within the specialized domain, and to have a grasp of the concepts and categories participating in them. In NATO, for example, the need for terminology management has long been recognized. This need is extensive to the armed forces of the countries within NATO.

This paper explains how the alphabetically organized (PD0-000) glossary of military terms of the Spanish Armed Forces (partially based on the *NATO Glossary of terms and definitions* (AAP-06)), was transformed into *MiliMarco*, a bilingual terminological knowledge base in which each concept appears within the context of a frame that highlights semantic relations and conceptual structure. Project objectives included the design and implementation of a frame-based military knowledge resource, information extraction techniques from a corpus of military documents, and the potential contribution to the interoperability of the resource with those of international organisms. Furthermore, a new objective has arisen during the presentation of the resource, which is the detection of inconsistencies in the Spanish military doctrine, since a concept-oriented frame-based approach can help to identify obsolete concepts and conceptual gaps within doctrinal documents.

The design and population of *MiliMarco* involved the analysis and transformation of the content in the glossary as well as the extraction of new information. For instance, specialized knowledge structures were extracted from the implicit structure of the definitions in the glossary and from the explicit lexicalization of semantic relations in the corpus. New concepts were added based on the gaps encountered in the glossary and new data categories were included, such as images, collocations and contexts.

The rest of this paper is organized as follows. Section 2 provides a concise outline of Frame-Based Terminology, the importance of context, and the lines of research initiated within this approach, as well as a brief account of previous efforts on military terminology management. Section 3 describes the PD0-000 military glossary and the method used to transform it into a knowledge base with an

underlying structure based on semantic relations. Section 4 presents the results and discusses some of the problems encountered. The conclusions are given in Section 5 as well as plans for future research.

2. Theoretical and Applied Framework

Specialized knowledge bases are based on explicit and implicit conceptual representations. Specialized knowledge understanding is enhanced when concepts and terms are organized so that the relations between them become explicit. Embedding concepts in a knowledge structure activates associative information in semantic memory and promotes context availability (Faber et al. 2014). Communication is facilitated by well-structured meanings that specify the relations between concepts as well as for situated or contextualized terminology. This is the main focus of Frame-Based Terminology (FBT) management (Faber 2012, Faber 2015) and the resources based on its principles.

Frame-Based Terminology is a theory that focuses on: (1) conceptual organization; (2) the multidimensional nature of terminological units; and (3) the extraction of semantic and syntactic information from multilingual corpora. The FBT approach to terminology and terminology management applies the notion of ‘frame’, defined as “a schematisation of experience (a knowledge structure), which is represented at the conceptual level and held in long-term memory and which relates elements and entities associated with a particular culturally embedded scene, situation or event from human experience” (Evans 2007, 85). Since frames highlight non-hierarchical as well as hierarchical conceptual relations, they provide a much richer representation.

Although the frame-like representations in FBT initially stem from Fillmore (1985, 222–254; 2006, 373–400; Fillmore et al. 2003, 298–332), they have been adapted to the structure of specialized knowledge units and their roles in specialized subject domains to include both language-specific and non-language-specific information. Frames are extracted from corpus texts in different languages through the use of knowledge patterns that encode semantic relations (Meyer 2001; Marshman 2006). The data thus obtained from the analysis of concordance lines are used to structure categories, create concept frames, and characterize general processes and actions. When frames are specified as an action or process with participants this provides a predicative frame linking two semantic categories. Although corpus data are used to extract information, the assumption is that the resulting frames encode conceptual knowledge that is non-language-specific. FBT thus focuses on how linguistic forms evoke or activate frame knowledge.

An FBT frame is thus a representation that integrates various ways of combining semantic generalizations about one category or a group of categories, whereas a ‘template’ is the representational pattern for individual members of the same category. In this way, frames become large-scale representations that link categories by means of semantic relations. They provide a basis for the selection of knowledge-rich linguistic, cultural, and graphical contexts. Until now, the only practical application of FBT has been EcoLexicon (ecolexicon.ugr.es), a multilingual terminological knowledge

base on environmental science, which includes 4471 concepts and 23,530 terms (Faber et al. 2016; San Martín et al. 2017). However, the methodology used to create EcoLexicon can be successfully implemented in other domains. The following sections describe how it was applied to military science in order to create a terminological knowledge base for the Spanish Armed Forces: *MiliMarco*.

With regards to previous efforts in the management of military terminology, NATO resources must be highlighted. NATO terminology is based on the Concise Oxford English Dictionary and Le Petit Robert. Specific NATO Agreed terminology is developed when the terminology contained in these dictionaries or that developed by recognized international standards organizations is inadequate for NATO purposes.

The general principles behind termhood and definitions are transparency, conciseness, stability, consistency, completeness and univocity. According to the NATO Terminology Directive, “the Alliance shall promote mutual understanding through the selection or development and use of commonly-agreed, well-defined, clear, precise, consistent and gender-neutral terminology, thereby enhancing the cohesion and effectiveness of the Alliance and its partner nations”.

Nevertheless, standardization is still far from ensuring efficient communication at all NATO settings. According to Jones (2015), language has been neglected in military history, despite the fact that conflicts are almost always between people who speak different languages. As an example, Jones and Askew (2014: 58) highlight the lack of reference resources that linguists had to face during the operation of Bosnia Herzegovina: “many of the linguists I met in SFOR⁶⁶ had therefore brought their own dictionaries to their offices. Not unsurprisingly, many different dictionaries were being used, which did not help to promote standardization of terminology”.

One possible reason for this could be the lack of interoperability of NATO glossaries as well as their format, since knowledge can only be accessed alphabetically. According to the policies in the NATO terminology Directive, terminology should be made available to the widest possible audience. For this reason, the new resource NATOTerm was created as the central repository for all non-classified NATO Agreed terminology in the near future.

NATOTerm concept-oriented and is thus structured in three levels, as is common practice in terminology management systems that are to be used in conjunction with CAT tools. There are different data categories at each level: (1) record level (security, domain, project, etc.); language level (approval status, definition, source, comments, notes, examples, related concepts, graphics, etc.); term level (type, source, acceptability, grammar, usage, approval status, etc.). As shall be seen in Section 3.1 *Milimarco* also follows the three-level structure with similar data categories.

⁶⁶ Stabilization Force, a NATO-led peacekeeping force after the Bosnian war.

Apart from terminology management, linguistic support in NATO covers both translation and interpreting (simultaneous, consecutive, and liaison), which may be required at a high level, provided by a qualified staff, or at a low level, provided by staff with more basic skills. Therefore, the users of NATO terminology include military linguists, civilian interpreters, editors, translators, assistants, and local personnel. The functions of linguistic support can be very diverse, such as command-level relations with authorities and parties, operations at the tactical and other levels, human intelligence, psychological operations, public affairs, legal affairs, contracting, logistics, policing, civil-military cooperation, administration of local personnel and training of indigenous forces, medical services, etc (NATO, 2011). The same target users and communicative situations can be considered for *MiliMarco*. Consequently, very diverse users need to gain specialized knowledge very quickly, since they may have to deal with a wide variety of subject fields within the same operation.

It is true that the former NTMS term base (NATO Terminology Management System) already included domain-related contextual information in certain entries by placing a qualifier at the beginning of a definition, but that was not enough. This is why NATOTerm is now being provided with conceptual structure in the form of a set of domains, known as the NATOTerm taxonomy (Jones 2011). These domains are mostly based on the range of subjects dealt with by the various NATO committees, agencies, and groups as well as on the documents they produce (i.e. political affairs, law and regulations, defence, etc.). This is in line with the frame-based approach. In this sense, as well as *MiliMarco*, NATOTerm is regarded as "not just a term-base, but a tool through which knowledge is shared"⁶⁷.

Thus, bearing in mind how NATOTerm was built (data categories and taxonomy) and who its targeted users and needs are, we built *MiliMarco* in a rather similar line. However, it is our claim that more meaningful access to knowledge can be provided by describing specialized concepts and terms in linguistically grounded structures such as frames.

3. Materials and Methods

As previously mentioned, the main aim of transforming alphabetically-ordered resources and making them more conceptually-based is to enhance knowledge acquisition by providing a more meaningful access to knowledge networks and frames. The official military glossary that required a 'total makeover' was the *PDO-000 Glosario de Términos Militares*, official publication of the *Mando de Adiestramiento y Doctrina* (MADOC) of the Spanish Armed Forces. It had initially been published in March 2004. As an ongoing project, it has undergone various revisions until the current glossary (2014), which has been adapted to the new doctrinal framework.

The current glossary is composed of 2,286 entries arranged in alphabetical order. Each entry is accompanied by a definition, though some of them (those related to polysemic terms) contain several definitions. The glossary is thus term-oriented, which required some adjustments during the conversion

⁶⁷ According to NatoTerm website (<https://nso.nato.int/natoterm/content/nato/pages/ntp.html?lg=en>)

process, as discussed in Section 4. Apart from definitions, term entries are sometimes accompanied by synonyms, variants, abbreviations and notes, which also needed to be reconfigured according to the concept-oriented approach.

The glossary uses different (sometimes confusing) means to express synonyms, variants and abbreviations. For example, when two terms can be used interchangeably, they both appear as the term entry, in italics and separated by a slash, as in the case of *autenticación/autenticación* (Figure 1). However, when two terms refer to the same concept but one of them is the preferred term, only the latter is included with the full description, whereas the non-preferred term is accompanied by the note *Véase* (See) followed by the name of the preferred entry (Figure 1). The same note (*Véase*) is also used when two entries are conceptually related. Nevertheless, when variants appear in the form of abbreviations they are followed by the term entry between brackets. Another use of notes is devoted to the inclusion of conceptual information pertaining to the concept entry (Figure 2), which in *MiliMarco* was reused in order to extract semantic relations (subsection 3.3).

The *Glosario de Términos Militares* (PD0 000) is a monolingual glossary with the exception of few entries that include the corresponding English term from the *NATO Glossary of Terms and Definitions* (AAP-06). This required merging existing English terms with their corresponding concept entry and the identification of missing equivalents. Nonetheless, there are certain equivalent terms that can be found within the definitions of NATO standardized terms (Figure 2).

Figures 1 and 2 show four entries activating the data categories of terms (with synonyms, variants, and abbreviations), definitions, English equivalents and notes. These data categories are a rather limited way of describing concepts. In *MiliMarco* they are reconfigured and expanded as shown in subsection 3.1.

*autenticación / autenticación*¹. Evidencia proporcionada por una firma, o un sello, de que un documento es auténtico y oficial.

AAP-06: *authentication*¹

*autenticación / autenticación*². Medida de seguridad destinada a proteger un sistema de telecomunicaciones contra las transmisiones fraudulentas.

AAP-06: *authentication*²

autodefensa. Véase “derecho de autodefensa”.

Figure 1. Three entries in the *Glosario de Términos Militares* (PD0 000)

fuerza de reacción (RF)². Una de las tres categorías de fuerzas de la Alianza (fuerza de reacción [RF: *reaction force*], fuerza principal de defensa [MDF: *main defence force*] y fuerza de aumento [AF: *augmentation force*]), versátil, muy móvil y capaz, mantenida a un alto nivel de operatividad.

NOTA: Están divididas en: fuerza de reacción inmediata (IRF: *immediate reaction force*) (más pequeñas) y fuerza de reacción rápida (RRF: *rapid reaction force*) (más capaces), ambas con componentes terrestres, navales y aéreos.

Figure 2. Entry containing English terms in the definition and conceptual information in the note

The subsections that follow explain the main steps in the process of transforming the glossary into a military knowledge base.

3.1. Terminology knowledge base design

Besides being valuable tools for storing information, a terminology knowledge base should be the practical application of an approach to terminology management. In fact, the creation of a termbase stems from an explicit (or implicit) commitment to a set of premises regarding knowledge representation. The transmission and acquisition of technical information is enhanced when knowledge resources are designed so that users, whether human or artificial, can easily access concepts and associate information in order to understand and acquire specialized knowledge. Evidently, the design parameters should be based on user needs as well as their reasons for consulting the knowledge resource. As previously stated, in *MiliMarco* we considered the same user types and communication needs as those considered in NATOTerm.

The microstructure of a terminological knowledge base comprises the data categories in each term entry. The selection of fields reflects the information that users wish to know about the concept (e.g. definition, part-of-speech, context, collocations, etc.). In regards to the macrostructure of the term base, the choice of structure is just as important since it affects the speed and types of possible knowledge access (e.g. frame-based, semantic networks, alphabetical order, etc.). From a multilingual perspective, when terminology resources are conceptually based, conceptual structure can act as an anchor point for linking terms in different languages. In this way, it also provides a foundation for interlinguistic correspondence.

The new design of the term entry in the *MiliMarco* [*MiliFrame*] knowledge base was partially based on NATOTerm as well as the recommendations of TerminOrgs (Terminology for Large Organizations), a consortium of terminology professionals that foment terminology management as part of the identity construction and communication strategy of large companies and organisms. From the TBX standard provided by TerminOrgs, we selected a series of data categories based on the latest ISO standards, which were tailored to the context of the Spanish Armed Forces. Table 1 summarizes the structure of

these fields at three levels (concept-language-term) along with the nature of each data category (i.e. obligatory, automatic, pick list, free text, etc.).

Level	Descriptive fields / Data categories		Type
	ID		Obligatory, automatic
	Date created		Obligatory, automatic
	Most recent modification		Obligatory, automatic
	Frame		Obligatory, pick list
	Conceptual category		Obligatory, pick list
	Semantic relations		Obligatory, pick list
	Image		Optional, multimedia file
Concept	Note		Optional, free text
		Definition	Obligatory, free text
Language	Spanish, English	Definition source	Obligatory, free text
Term	Terms or synonyms	Part-of-speech	Obligatory, pick list
		Term type	Optional, pick list
		Context	Optional, free text
		Context source	Optional, free text
		Collocations	Optional, free text
		Note	Optional, free text

Table 1. Knowledge base structure

These data fields provide the contextual information for each entry in the form of usage examples (i.e. context) and phraseological constructions (i.e. collocations). The design of the resource also accounts for the theoretical and practical considerations in NATO documents on terminology management (e.g. *Guidance for the Development and Publication of NATO Terminology*). For example *Term type* states whether the term is obsolete or preferred, and whether it is an abbreviation, acronym, etc. *Semantic relations* associates each entry to others that are semantically linked to it, similar to the field “related terms” in NATOTerm.

3.2. Identification of inventory of basic semantic categories and relations in the glossary definitions

Although information was extracted from linguistic data, the assumption was that the resulting frames encoded conceptual knowledge that was non-language-specific. However, non-language specific information not only comes in the form of semantic relations, but also in the form of conceptual invariants encoded in a wide range of languages that are used for specialized communication.

An in-depth analysis of the glossary can reveal the underlying structure of the domain. Any glossary of specialized knowledge units tells a story about the domain and the contexts activated within it. The terms and their patterns are like pieces in a jigsaw puzzle. The conceptual structure underlying the glossary can be extracted by specifying the relations between terms and then filling in the empty spaces. The terms in the glossary evidently encode the important actions and processes carried out, the actors or agents that participate in them, and the instruments used to perform them. The most salient frames or the knowledge structures that link categories and concepts are indicative of the most prototypical actions, processes, and events that take place within the domain.

For design purposes, language structure was used as a conceptual mirror to extract the structure of a domain from the terminographic definitions in the glossary. Firstly, the superordinate term (*genus*) in each definition was used as a guideline for assigning each concept a general category. Then, semantic relations were extracted from the definitions' *differentiae* in order to relate categories in a general frame-like structure and concepts in semantic networks (3.4). Thus, the glossary was first converted into a pre-network structure derived from the glossary's definitions and then enriched with corpus information (see 3.2).

For example, from the definitions in Table 2 we can infer that BLISTER AGENT and PULMONARY AGENT are both hyponyms of CHEMICAL WEAPON AGENT, since both are defined as CHEMICAL AGENT whereas CHEMICAL AGENT contains a more superordinate *genus* (i.e. CHEMICAL SUBSTANCE). Therefore, the genus makes category membership explicit.

Furthermore, from the analysis of *differentiae* we can extract the typical roles participating in military events (i.e. AGENT, ACTION, PATIENT, RESULT. etc.) as well as the semantic relations (i.e. *type_of*,

used_during, has_function, affects, result_of, causes, etc.) according to which the domain can be structured.

Agente químico de guerra: sustancia química que puede ser empleada en operaciones militares para matar, herir gravemente o incapacitar al personal mediante sus efectos fisiopatológicos.

*Translation*⁶⁸

Chemical weapon agent: chemical substance [AGENT/*type_of*] that can be used in military operations [ACTION/*used_during*] to kill, injure or incapacitate [ACTION/*has_function*] man [PATIENT/*affects*] through its physiopathological effects [RESULT/*results_in*].

Agente vesicante/dermotóxico: agente químico que produce, en contacto con la piel, lesiones similares a las quemaduras. También puede producir efectos en los ojos y en el tracto respiratorio si es inhalado.

Translation

Blister/vesicant agent: chemical agent [AGENT/*type_of*] that produces burn-like injuries [RESULT/*causes*] when in contact with the skin [PATIENT/*affects*]. It can also affect eyes and the respiratory tract [PATIENT/*affects*] if inhaled [ACTION/*means*].

Agente sofocante/neumotóxico: agente químico que afecta, fundamentalmente, al aparato respiratorio al ser inhalado.

Translation

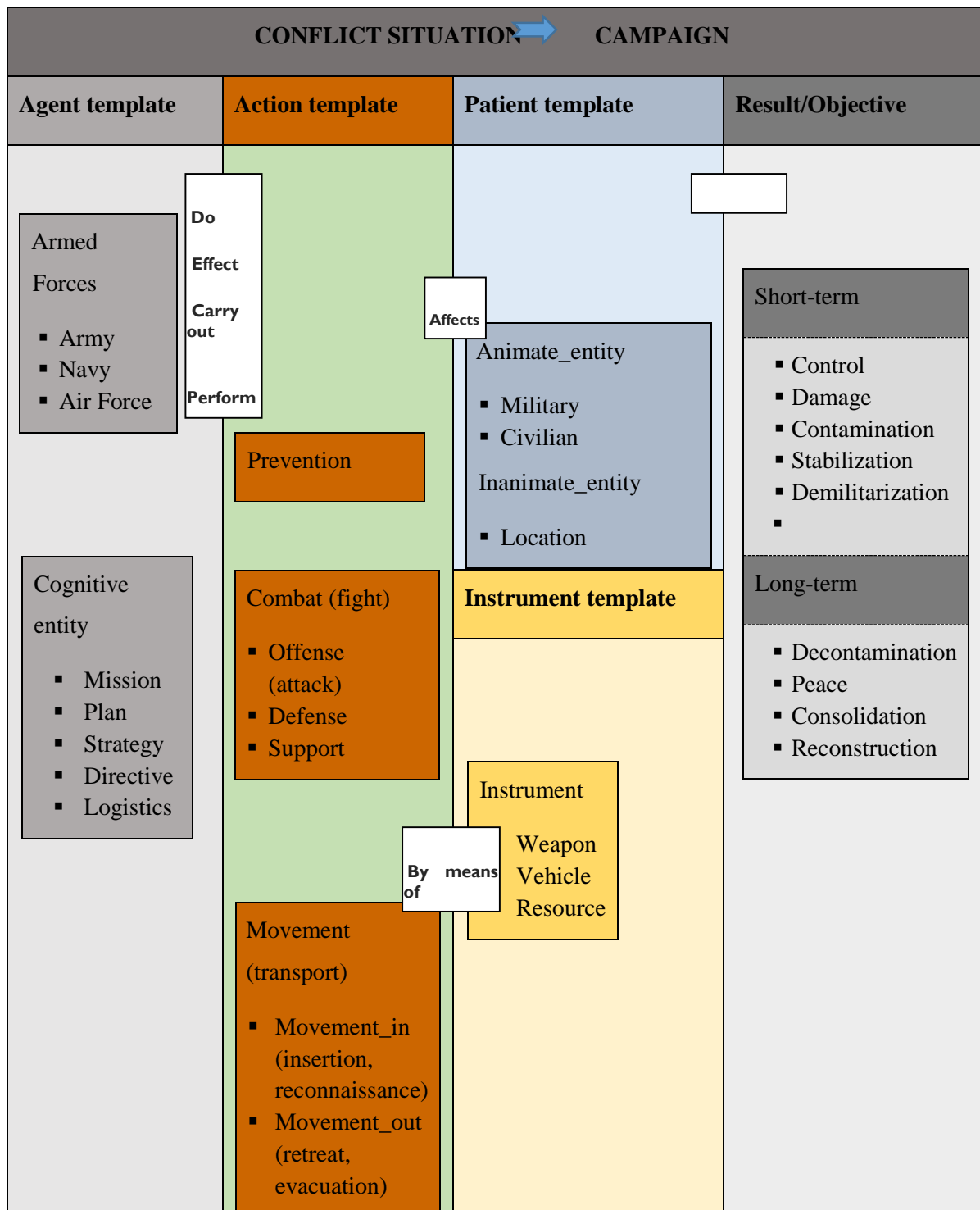
Pulmonary/choking agent: chemical agent [AGENT/*type_of*] that affects the respiratory system [PATIENT/*affects*] when inhaled [ACTION/*means*].

Table 2. Definitions of CHEMICAL AGENT and its hyponyms BLISTER AGENT and PULMONARY AGENT in the glossary

All of the glossary entries were analyzed following this procedure and classified in an inventory of basic categories, namely entities, actions, time, space, and attributes (see appendix). Entities are divided into ANIMATE_ENTITY and INANIMATE_ENTITY. INANIMATE_ENTITY is subdivided into CONCRETE and ABSTRACT. There are also general categories for ACTION, SITUATION, MEASUREMENT, and ATTRIBUTE. The main categories within ANIMATE_ENTITY are MILITARY ROLE, MILITARY GROUP, INSTALLATION, and EQUIPMENT. Important abstract entities are cognitive (PLAN, STRATEGY) and regulatory (RULES, REGULATIONS, PRINCIPLES). In regard to actions, not surprisingly, the most important are those related to COMBAT, MOVEMENT, DEFENSE/PROTECTION, and MANIPULATION (especially use of WEAPONS and VEHICLES). Finally, the most important types of attribute are those related to CAPACITY and POWER.

⁶⁸ Definition translations are provided for the sake of clarity in this paper, but the knowledge base only includes Spanish definitions.

These semantic categories hold relations with each other in different military frames composed of the following prototypical participants: AGENT, ACTION, PATIENT and RESULT/OBJECTIVE. These are the same participants that could be extracted from the analysis of definitions in Table 2. The basic template shown in Figure 3 is the general military event where all concepts and categories can accommodate within the domain. Agents are usually the armed forces (e.g. NAVY) or mental entities (e.g. PLAN) that initiate an action (e.g. attack) with an instrument (e.g. WEAPON) affecting patients, which can be animate entities (e.g. CIVILIAN) or locations (e.g. COUNTRY), causing a result (e.g. PEACE).



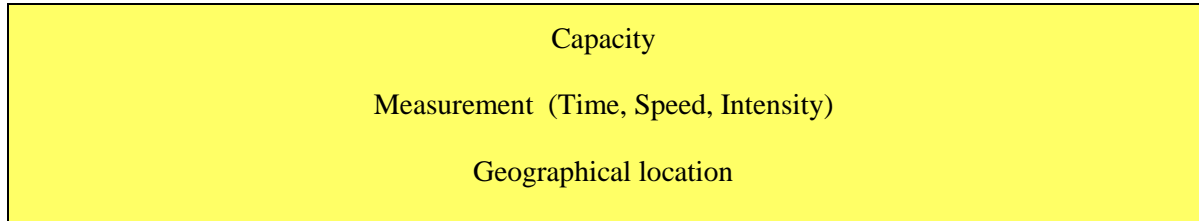


Figure 3. General military event

3.2. Corpus compilation and information extraction

As previously mentioned, FBT frames are extracted from corpus texts in different languages through the use of knowledge patterns that encode semantic relations (Meyer 2001; Marshman 2006; León-Araúz and Faber 2016; León-Araúz and San Martín 2018). Knowledge pattern (KP) queries in the form of sketch grammars created by León-Araúz, San Martín, and Faber 2016) extract concordance lines where the term in question appears related to others. The data thus obtained are used to structure categories to create concept frames as well as to characterize general processes and actions. Therefore, corpus information complements the categories and relations firstly extracted from the definitions in the glossary.

A bilingual corpus of approximately 30,000,000 words was compiled in English (15 million words) and Spanish (15 million words). This exhaustive documentation process required the collaboration of various branches of the Spanish Armed Forces as well as access to classified information.

The corpus compilation process was composed of the following stages: (i) identification of relevant documents; (ii) downloading files and converting them into txt format; (iii) cleaning of files to avoid codification problems; (iv) uploading files to a corpus analysis application (i.e. Sketch Engine); (v) corpus compilation with a lemmatizer, a POS Tagger and the KP-based sketch grammars.

The compilation of both corpora was carried out both manually and automatically. Approximately half of the documents were those provided by the MADOC, which came from sources such as *Revista Española de Defensa*, *Boina Negra*, *Revista de Sanidad Militar*, publications on military doctrine, instruction manuals, as well as a wide range of NATO documents. When both corpora were uploaded to Sketch Engine, terms were automatically extracted in order to obtain terms which, though not included in the glossary, were sufficiently representatives to be included in the knowledge base. This also provided us with seed words, which could be used to automatically search Internet for new corpus texts with the WebBootCaT tool, integrated in Sketch Engine. These key words also included a selection of those from the glossary so that the corpus texts would be in consonance with the terms in the glossary. This allowed us to double the size of the corpus.

Once the corpus was compiled, the following types of data categories were extracted from the corpus and populated in the knowledge base: frames and conceptual categories, semantic relations, synonyms, equivalents, contexts and collocations. Sketch Engine offers different functionalities that can assist in

the extraction of such data. For example, bilingual word sketches help in the establishment of bilingual correspondences as well as on the extraction of semantic relations, contexts, collocations and new concepts and terms.

An example is provided in Figure 4. The modifiers of both *operación* and *operation* provide many hyponyms of the concept, whereas the verbal sketches provide interesting information for the extraction of both semantic relations and collocations. For instance, the verbs that collocate with *operación/operation* as an object (e.g. *conducir* and *conduct*; *ejecutar* and *execute*; *apoyar* and *support*) can be extracted in a parallel view and their equivalence can be easily established.

operación (noun) Milimarco_ES freq = 27,708 (1,583.35 per million) **operation** Milimarco_EN freq = 20,866 (3,338.93 per million)

Use another candidate translation: [Operation](#) [Atalanta](#) [military](#) [Chad](#) [Congo](#) [ESDP](#) [deployment](#) [exercise](#) [NATO](#)

Click on collocates to access reciprocal bilingual search or find [translated collocations](#)

modifiers of "operación"		modifiers of "operation"		verbs with "operación" as object		verbs with "operation" as object	
28.31		74.70		17.76		19.52	
especial	576 10.50	military	1,264 10.52	conducir	219 10.31	conduct	759 11.80
militar	1,921 10.46	joint	1,012 10.42	realizar	443 10.05	support	358 10.32
anfíbio	239 9.78	combat	411 9.46	ejecutar	209 9.97	execute	149 9.95
ofensivo	234 9.71	air	389 9.12	apoyar	176 9.72	plan	93 9.26
conjunto	272 9.58	major	294 9.10	dirigir	134 9.19	sustain	71 8.96
táctico	285 9.41	multinational	287 9.08	desarrollar	173 9.09	coordinate	77 8.82
psicológico	139 9.01	special	246 8.81	combinar	86 8.87	affect	65 8.70
aéreo	310 8.83	stability	221 8.80	sostener	79 8.78	integrate	61 8.68
bélico	108 8.62	nato	272 8.71	iniciar	83 8.56	peacekeepe	50 8.60
naval	137 8.55	contingency	194 8.60	efectuar	78 8.56	synchronize	48 8.46
defensivo	104 8.49	future	193 8.52	llevar	81 8.54	deploy	55 8.35
futuro	95 8.33	maritime	170 8.34	aerotransportada	56 8.51	direct	44 8.24
multinacional	86 8.29	cyberspace	161 8.34	lanzar	61 8.36	shape	40 8.16
terrestre	113 8.24	support	176 8.18	aerotransportadas	47 8.24	facilitate	44 8.11
diferente	102 8.11	amphibious	138 8.13	planear	49 8.22	enable	44 8.03
convencional	61 7.75	evacuation	149 8.03	liderar	48 8.13	control	39 8.03
duradero	48 7.59	cyber	149 7.93	comenzar	49 8.11	include	74 7.73
retrogrado	45 7.53	information	137 7.87	encubrir	43 8.06	perform	34 7.73
actual	70 7.53	offensive	114 7.86	permitir	62 8.04	continue	27 7.61
civil	94 7.48	medical	195 7.85	coordinar	49 7.97	design	27 7.50
específico	57 7.22	land	118 7.85	relacionar	64 7.94	characterize	24 7.49
protector	37 7.19	current	103 7.61	decir	74 7.70	base	30 7.44
posible	49 7.12	combined	94 7.57	incluir	65 7.68	lead	25 7.42
internacional	90 7.02	other	141 7.50	determinar	70 7.64	be	151 7.33
aeroterrestre	32 7.02	relief	88 7.50	emprender	32 7.59	fly	20 7.26

Figure 4. Bilingual word sketch of *operación-operation* in Sketch Engine

In contrast, other verbs such as *incluir* and *include* can be used to extract semantic relations and/or knowledge-rich contexts, since they often act as KPs expressing hyponymic or meronymic relations, as shown in Figure 5. For example from the concordances in Figure 5, *nuclear war* can be extracted as a hyponym of *military operation*, whereas *contaminated operations* and *uncontaminated operations* can be divided into different phases: *triage* and *emergency treatment*, and *treatment* and *final disposition*, respectively.

and lacks logistic support. Airborne **operations include** the following: a. Paratroop. These patient patterns. Due to their nature these **operations include** a large portion of medical support. In FHP measures during all phases of the **operation including** pre- and post-deployment. (5) spectrum of actions provided during such **operations includes** : (1) All actions conducted by during the envisaged spectrum of military **operations including** nuclear war. Basic nuclear and that may be encountered in military **operations including** those deriving from a nuclear This will include details on location, hours of **operation including** sick parade/walk-in emergency to all management aspects of a food services **operation including** food safety risk. To ensure that the terrain. Some recent examples of urban **operations include** Mogadishu, Sarajevo, Baghdad, for conducting medical evacuation in arctic **operations include** the following: Arctic warfare is) and uncontaminated (clean). Contaminated **operations include** triage, emergency treatment, and ind patient decontamination. Uncontaminated **operations include** treatment and final disposition. All and space weather operations. These **operations include** collecting, analyzing and predicting principles, to restore or maintain peace. Such **operations include** conflict prevention, peace 50 See Contribution to Joint Operations. 5 Maritime **operations include** any actions performed by surface, covers all activities prior to arrival on **operations including** : warning; reconnaissance; to detection and attack. Counter-air **operations include** all actions, taken by any component, to force operations involving SOF to follow-on **operations including** conventional forces. This ensures to detection and attack. Counter-air **operations also include** all actions, taken by any

5.1.1. Overview 1. Space support to **operations includes** all activities that provide

Figure 5. Verb *include* as a KP in the English corpus.

From all of the entries included in the glossary, a first set was selected and queried in the corpus based on the multiword terms that contained a common head. In this way, conceptual gaps in the glossary were easily identified and new concepts were rapidly accommodated in the conceptual structure derived from the glossary.

For instance, there are multiple terms that contain *operación* [*operation*] as their head (e.g. *operación especial, anfibia, de mantenimiento de la paz, decisiva, de intervención limitada, de búsqueda y rescate, de extracción, de evacuación de no combatientes*, etc.). However, not all of the compounds reflected in the corpus were included in the glossary, such as *operación táctica*, in spite of being among the first results of the modifiers of *operación*.

From the concordances of *operación táctica* (Figure 6), which can be reused as contexts in the knowledge base, other types of information can be extracted, such as hyponyms (e.g. *operación psicológica táctica* or *operación aéreo táctica*), other related concepts (e.g. *centro de operaciones tácticas*), synonyms and morphosyntactic variants (e.g. *operación aérea táctica* and *operación aéreo táctica*) abbreviations (e.g. OPSIC for *operación psicológica táctica*) and collocations (e.g. *llevar a cabo, ejecutar, realizar*, etc.).

no se limitan a un área geográfica restringida.	OPERACIONES PSICOLÓGICAS TÁCTICAS : Son las OPSIC que se
Teatro de Guerra principal donde se lleva a cabo	operaciones tácticas o estratégicas. etc. el efecto
. Una fuerza aérea encargada de llevar a cabo	operaciones aéreo tácticas en coordinación con las fuerzas
(EAF)] que trabajan dentro de cada centro de	operaciones tácticas . es preciso proteger y evitar que
como una porción funcional dentro del centro de	operaciones tácticas de la fuerza cuando unidades de
Normalmente se encuentra junto con el centro de	operaciones tácticas del cuerpo de ejército. CENTRO DE
Normalmente se encuentra junto con el centro de	operaciones tácticas del cuerpo de ejército. CENTRO DE
del ejército que trabaja en los centros de	operaciones tácticas (COT) a los niveles de cuerpo de
. En Marina. Pueden ser objetivos locales de	operaciones tácticas . que sirve para guardar y/o llevar
de tal forma que concuerde con la	operación táctica esperada de la organización embarcada
una información oportuna acerca de las	operaciones tácticas terrestres de combate. </p><p> 718 </p>
de servicio de aeronaves y de alerta aérea de las	operaciones aéreo tácticas en un área de responsabilidad.
de servicio de aeronaves y de alerta aérea de las	operaciones aéreo tácticas en un área de responsabilidad.
el control del movimiento y la conducción de las	operaciones tácticas en las que participan tropas bajo su
el control del movimiento y la conducción de las	operaciones tácticas en las que participan tropas bajo su
AÉREO OFENSIVO: (AAO) Aquella parte de las	operaciones aéreas tácticas en respaldo directo de las
de estado mayor a planear y ejecutar las	operaciones tácticas sin requerir la presencia y
asignadas al mismo teatro a fin de realizar las	operaciones aéreas tácticas previstas. FUERZA AÉREA
durante el planeamiento o la ejecución de una	operación táctica . El efecto de las operaciones tácticas
adiestrar a los jefes a planear y a ejecutar una	operación táctica y para demostrar la ejecución de una

Figure 6. Concordances of *operación táctica*

3.4. Formulation of semantic networks based on the analysis of conceptual propositions

Starting from the basic structure of the general military event (subsection 3.2), and based on the analysis of both definitions and corpus information, concepts were structured in semantic networks through the identification and representation of semantic relations, as shown in Figure 7 for the concept OPERACIÓN MILITAR [*military operation*]. The inventory of semantic relations so far is the following: *type_of*, *part_of*, *phase_of*, *instrument_of*, *controls*, *location_of*, *attribute_of*, *target_of*, *affects*, *domain_of*, *effected_by*, *represents*, *measures*, *destroys*, *delimited_by*, *result_of*, *causes*, *has_function*.

In Figure 7 OPERACIÓN MILITAR is mostly related through the *type_of* relation to its hyponyms, although meronymic relations (MISIÓN *part_of* OPERACIÓN MILITAR) and non-hierarchical relations (MARCO OPERATIVO *representa* OPERACIÓN MILITAR) also appear.

Definitional hierarchy	
	military operation: set of military actions coordinated in time, space, and purpose to achieve a military objective at level of tactical, operational, or strategic leadership as established in a directive, plan, or order.
➔	main operation: military operation involving the coordinated action of large forces in a phase of a campaign to achieve operational objectives.
➔	support operation: main operation whose objective is to create and maintain the fighting capacity of operational organizations and which ensure the necessary capacities to perform other operations.
➔	peace support operation: support operation consistent in military operations, which under the auspices of the UN or another international organization has the purpose of supporting and fomenting diplomatic efforts and political processes to avoid, contain, moderate or resolve conflicts.

Table 3. Definitional hierarchy of military operation and subtypes

Since the definitions only provided very general information, it was necessary to enrich them with corpus data in the form of other multi-word terms, whose structure reflected that of the underlying semantic frame. For example, MILITARY OPERATION reflects an agent slot, whereas PEACE SUPPORT OPERATION reflects the purpose of the operation. The structure of the category of MILITARY OPERATION was found to have a five-dimensional structure, as specified as follows in the MWTs:

- AGENT slot: armed forces of one or various nations
- PURPOSE/OBJECTIVE slot: support (mainly for peace) and extraction (evacuation)
- LOCATION slot: air/water/land
- SCOPE slot: range/nature
- THEME slot: intelligence

Table 4 shows the full specification of each dimension based on the naming devices of the terms in Spanish. This means that the hyponymic relation *type_of* can be further specified according to each dimension and that the semantic network of OPERACIÓN MILITAR and its subtypes should be enriched with more non-hierarchical relations, namely those related to agents (e.g. *causes*, *result_of*), purpose (e.g. *has_function*), location (e.g. *delimited_by*, *location_of*), scope and theme (e.g. *has_function*, *affects*).

Dimension 1	Agent	Spanish term
Armed forces [General] [One nation]	military	<i>operación militar</i>
Armed forces [one nation]	armed forces (army, navy, airforce), paramilitary groups	<i>operación terrestre</i> <i>operación naval</i> <i>operación aérea</i> <i>operación de contraguerrillas</i>
Armed forces [More than one nation]	joint, multinational combined,	<i>operación conjunta</i> <i>operación combinada</i> <i>operación multinacional</i>
Dimension 2	Purpose/Objective	Spanish term
Support	support operation	<i>operación de apoyo</i>
	peace support <ul style="list-style-type: none"> • peace implementation • peace enforcement • peacekeeping • peace consolidation 	<i>apoyo a la paz</i> <ul style="list-style-type: none"> • <i>establecimiento de la paz</i> • <i>imposición de la paz</i> • <i>mantenimiento de la paz</i> • <i>consolidation de la paz</i>
	civilian groups	<i>apoyo a autoridades civiles en terreno nacional</i>
Extraction	extraction <ul style="list-style-type: none"> • military force • civilians 	<i>operación de extracción</i> <i>operación de evacuación de no combatientes</i>
Dimension 3	Location	Spanish term
Geographic sphere		<i>operación aeromóvil</i>
	Air	<i>operación aerotransportada</i> <i>operación de asalto aérea</i> <i>operación contra-aérea</i>
	Water	<i>operación anfibia</i>
	Land	<i>operación terrestre</i>
Dimension 4	Scope	Spanish term

Range	Range	<ul style="list-style-type: none"> • <i>operación de intervención limitada</i>
	<ul style="list-style-type: none"> • limited • open • specific 	<ul style="list-style-type: none"> • <i>operación abierta</i> • <i>operación específica</i>
Nature	Nature	<ul style="list-style-type: none"> • <i>operación psicológica</i> <ul style="list-style-type: none"> ◦ <i>operación de información</i>
	<ul style="list-style-type: none"> • psychological <ul style="list-style-type: none"> ◦ information • clandestine • humanitarian 	<ul style="list-style-type: none"> • <i>operación clandestina</i> • <i>operación humanitaria</i>
Dimension 5	Theme	Spanish term
Intelligence	<ul style="list-style-type: none"> • information sources 	<ul style="list-style-type: none"> • <i>operación de fuentes</i>

Table 4. Specification of the five-dimensional structure of OPERACIÓN MILITAR

Results and Discussion

This project posed a series of challenges related to the existing as well as the missing information in the glossary. For example, the definitions, many of which were not consistently formulated or did not adequately clarify the term defined, could not be altered to make their structure more uniform or clearer, as they are standardized by the Spanish military forces.

Another challenge stems from the conversion of a term-oriented glossary into a concept-oriented knowledge base. This meant creating different entries for those where different definitions were provided; disambiguating polysemic terms; adding synonyms and abbreviations (as monolingual variants) as well as English equivalents at the language level of each entry; spotting different definitions that actually pointed to the same concept (sometimes several definitions were provided not because the term was polysemic but because the sources were different); and converting the semantic information contained in the definitions and notes of each entry into conceptual propositions in the knowledge base.

The few English terms contained in the glossary were linked to each concept entry and new equivalents were included for the rest of them, based on other resources (especially NATOTerm where possible) and corpus information. In addition, new concepts and terms were included based on the gaps encountered in the glossary and the terms extracted from the corpus.

After applying the automatic term extraction functionality in Sketch Engine, a list of term candidates was collected. These new terms were added (1) as synonyms and variants in existing entries, (2) as new concept entries when they were not available in the glossary or (3) as collocational information associated with terms.

New concept entries were also created based on the analysis of existing multi-word expressions (MWEs) and querying the corpus with their most recurrent heads. For example, Table 5 shows all concepts related to ACTION, those included in the glossary and new additions from the corpus.

Head	Hyponyms included in the glossary	Hyponyms extracted from the corpus
Action	acción conjunta, acción de apoyo, acción de apoyo a autoridades civiles, acción de conjunto, acción de conjunto-refuerzo, acción de estabilización, acción de fuego , acción de fuego tipo, acción defensiva, acción directa, acción en profundidad, acción envolvente, acción fijante, acción frontal, acción lateral, acción lejana, acción libre, acción militar táctica, acción ofensiva, acción prohibida, acción restringida, acción retrógrada	acción exterior, acción enemiga exterior, acción común, acción conjunta, acción terrorista, acción hostil, acción táctica, acción estratégica, acción de guerra, acción defensiva directa, acción específica directa, acción exterior comunitaria, acción enemiga exterior, acción exterior europea, acción ofensiva contundente, acción ofensiva enemiga, acción ofensiva terrestre, acción aérea ofensiva, acción militar conjunta, acción militar táctica, acción bilateral conjunta, acción directa preventiva, acción cívica militar, acción táctica complementaria

Table 5. Concept expansion based on the analysis of recurrent heads in MWEs

Regarding semantic networks, thanks to definition and MWE analysis, word sketches and KP-based sketch grammars and queries, semantic relations were extracted from the corpus and represented in the knowledge base in the form of conceptual propositions. Finally, images were collected from public repositories according to the information conveyed and the nature of the concepts described, trying to enhance the conceptual structure shown in their semantic network.

Currently *MiliMarco* shows all this information in a dynamic concept-oriented interface as shown in Figure 8.

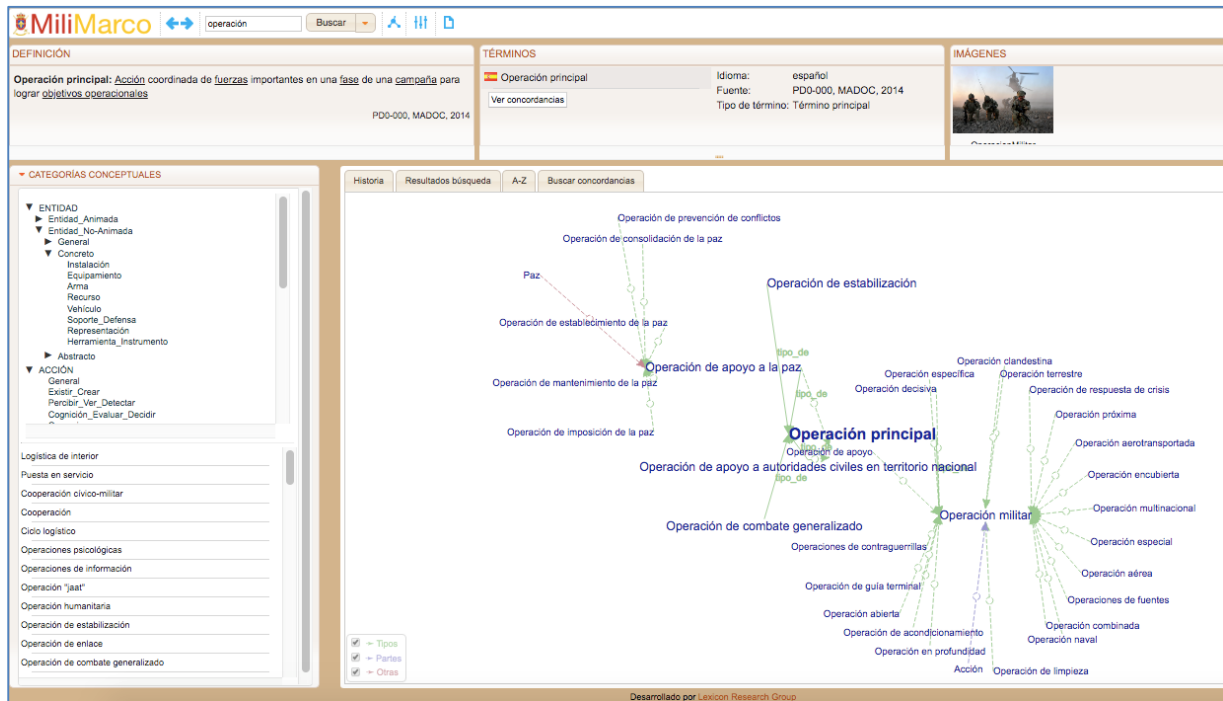


Figure 8. MiliMarco user interface

As can be seen, the main area shows the conceptual representation of military notions. On the left panel, users can access entries browsing through conceptual categories, whereas on the right and main area of the screen, semantic networks are shown. Users can click on any of the concepts in this network and reconfigure their structure around the new concept in an interactive way. They can change the settings to customize the networks (i.e. number of nodes, relations displayed, distance between nodes, labels of the relations, etc.).

There are also different access and visualization modes of concepts, such as a tree-like structure (Figure 9), where only *type_of* relations are shown, and an alphabetical access (Figure 10).

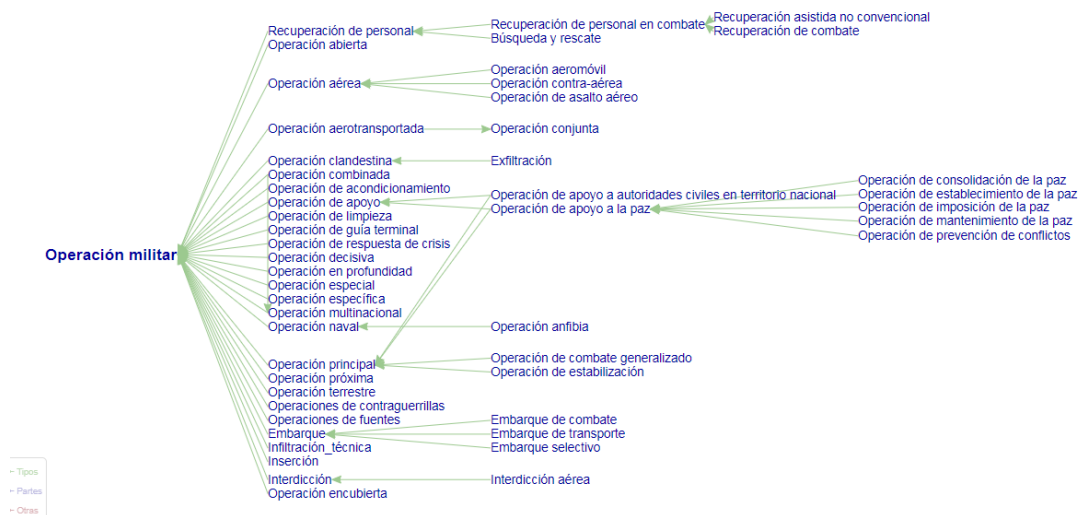


Figure 9. Tree-like structure of OPERACIÓN MILITAR



Figure 10. Alphabetical access to concept entries

In the upper ribbon, users can find: (1) definitions, where all terms included in the knowledge base have a hyperlink to their entry; (2) term information, such as terms, synonyms, variants, equivalents, sources, term types, etc. and access to concordances; (3) images. When clicking in the *Ver concordancias* button or when typing the term in the corresponding box, different contexts from the corpus are shown (Figure 11).



Figure 11. Term box and access to the corpus

4. Conclusions

Inevitably, the design and information included in a terminology knowledge base depend on user needs and the decoding and/or encoding tasks to be carried out by them. This is less of a question of the number of data categories, and more of a question of effective information access, extraction, and analysis. Conceptual data can be extracted from definitions (semantic analysis) and texts (corpus analysis), or by eliciting information from experts by means of a protocol involving a questionnaire, discussion group, a series of intensive interviews, etc. The designer can also use both methods and place his/her main focus on one or the other, depending on the context.

A terminological knowledge base can have an alphabetical search mechanism, but at the same time, it can also allow users to opt for a conceptual search. Although electronic resources do not have the space constraints of paper documents, other problems can arise since decoding the meaning underlying a set of terms in a specialized field is an extremely complex task. One of the difficulties is the perception of conceptual similarities, relations, and patterns in term meaning within a highly specialized domain such as military science. The other difficulty is linked to the choice and configuration of design parameters.

For this proposal, we took the *Glosario de Términos Militares* (PD0-000) and structured it conceptually. Our analysis was mainly based on semantic and corpus analysis though the results were also subjected to expert validation. The implicit conceptual structure underlying the glossary highlights the important structuring role of actions and processes in regards to object categories, as shown in the general military event and as exemplified in the multidimensional structure of military operations.

References

- Cabezas-García, M. & Faber, P. (2018) Phraseology in specialized resources: an approach to complex nominals. *Lexicography*, 5(1):55-83
- Evans 2007
- Faber, Pamela (ed.) 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: De Gruyter.
- Faber, P., Verdejo, J., León-Araúz, P.; Reimerink, A. & Guzmán, G. 2014. Neural Substrates of Specialized Knowledge Representation. *Revue française de linguistique appliquée* 19 (1): 15-32.
- Faber, P. 2015. Frames as a Framework for Terminology. In: *Handbook of Terminology*, H.J. Kockaert, & F. Steurs (eds.). Amsterdam/Philadelphia: John Benjamins. 14-33.
- Faber, P., León-Araúz, P. & Reimerink, A. (2016) EcoLexicon: new features and challenges. In *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, edited by Kernerman, I., Kosem Trojina, I., Krek, S. & Trap-Jensen, L., pages 73-80. Portorož
- Fernández-Domínguez, J. 2016. A morphosemantic investigation of term formation processes in English and Spanish. *Languages in Contrast* 16 (1). 54–83.

- Fillmore, C. J. 1982. Frame Semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*, 111-137. Seoul: Hanshin.
- Fillmore, C. J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6 (2). 222-254.
- Fillmore, C. J. 2006. Frame Semantics. In Dirk Geeraerts (ed.), *Cognitive Linguistics. Basic readings*, 373-400. Berlin and Boston: De Gruyter.
- Fillmore, C. J., Johnson, C.R. and Petruck, M.R.L. 2003. Background to FrameNet. *International Journal of Lexicography* 16 (3). 235-250.
- Gallese, V. and Lakoff, G. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology* 22 (3-4). 455-479.
- Jones, I. 2011. The NATO Terminology Programme and NATOTerm. *33rd Translating and the Computer Conference*. London, 17-18 November 2011.
- Jones, I. and Askew, L. 2014. *Meeting the language challenges of NATO operations: policy, practice and professionalization*. London/New York: Palgrave MacMillan.
- Jones, I. 2015. *Book Talk: Meeting the language challenges of NATO operations*. 24 March 2015. Available at: http://www.natolibguides.info/library/booktalk_jones
- NATO. 2011. Linguistic Support for Operations, *ALING P-1*. Available at: <http://nso.nato.int/nso/zPublic/ap/alingp-1.pdf>
- León-Araúz, P. & San Martín, A. (2018) The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, edited by Kerneman, I. & Krek, S., pages 94-99. Miyazaki: Globalex
- León-Araúz, P., San Martín, A. & Faber, P. (2016) Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 73-82. Osaka, Japan: COLING 2016.
- Marshman, E. 2002. The cause-effect relation in a biopharmaceutical corpus: English knowledge patterns. In: *Proceedings of the 6th international Conference on Terminology and Knowledge Engineering*. Nancy (France). 89-94.
- Meyer, I. 2001. Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework. In: *Recent Advances in Computational Terminology*. Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds.). Amsterdam: John Benjamins. 279-302.
- Sager, Juan C., Dungworth, David and Peter F. McDonald. 1980. *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

San Martín, A., Cabezas-García, M., Buendía, M., Sánchez-Cárdenas, B., León-Araúz, P. & Faber, P. (2017) Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America*, 38(1):96-115.

Štekauer, Pavol, Valera, Salvador and Livia Körtvélyessy. 2012. *Word-formation in the world's languages: a typological survey*. Cambridge: Cambridge University Press.

Appendix

• ENTITY

- ▼ Animate Entity
 - Military/Social Role
 - ▼ Group_Organization
 - Military Group
 - Civilian Group
- ▼ Inanimate entity
 - ▼ General Inanimate Entity
 - Configuration_System
 - ▼ Concrete Inanimate Entity
 - Installation
 - Equipment
 - Weapon
 - Resource
 - Vehicle
 - Support_Defense
 - Representation
 - Tool_Instrument
 - ▼ Abstract Inanimate Entity
 - ▼ Cognition
 - Plan
 - Strategy
 - Method
 - Judgement_Decision
 - ▼ Communication
 - Utterance
 - Document

- Sign
 - Information
 - Control_Regulation
- ▼ ACTION
 - General
 - Exist_Create
 - Perceive_See_Detect
 - Think_Evaluate_Decide
 - Communicate
 - ▼ Change_Transform
 - Repair_Maintain
 - Move
 - ▼ Use_Manipulate
 - Use_weapon
 - Use_vehicle
 - Fight_Combat
 - Protect_Defend
 - ▼ Possess_Obtain
 - Give_Supply
 - Receive
 - Transfer
 - Give_Receive:Treatment
 - Obtain_Information
 - Control
 - Locate_Position_Organize
 - Work_Function
 - Construct
- ▼ SITUATION
 - Condition
 - ▼ State
 - Conflict
 - ▼ Space
 - Trajectory_Path
 - Position_Location_Zone

- ▼ MEASUREMENT
 - Time_Phase
 - Vertical_Horizontal_Distance
 - Volume_Number_Quantity
 - Intensity_Force
 - Point_Angle
 - Limit
 - Level_Degree
 - Cost
- ▼ ATTRIBUTE
 - Power
 - Capacity

THE NAME 'NDEBELE' CAN SUGGEST THE SAME OR A DIFFERENT LINGUISTIC GROUP

Dr K.S Mahlangu, iZiko lesiHlathululi-mezwi sesiNdebele,

University of Pretoria

Abstract

In Africa as a whole, amaNdebele are the Nguni linguistic groups that are found in the Republic of South Africa and Zimbabwe. The amaNdebele of the South, i.e. the Southern and Northern Ndebele and Zimbabwe Ndebele are distinguished mainly on the basis of their historical, traditional and cultural backgrounds. Some language scholars sometimes confuse the two amaNdebele groups as a result of having no linguistic evidence or even been made to support the theories propounded by the historians and anthropologists in regard to the distinctiveness of the two language groups. The fact that these groups share the same title names, “amaNdebele” referring to the nation and isiNdebele referring to the languages that they speak, necessitates the need for investigation. In the Republic of South Africa, “isiNdebele” is one of the official languages which is spoken by amaNdebele who claim their origin from Chief Musi. The same linguistic name, ‘isiNdebele’ is a language spoken by the followers of Mzilikazi who are mainly spread and found in the southern region of Zimbabwe. SiNdebele which is a language spoken by amaNdebele who are the followers of Gegana, i.e. the Northern Ndebele, is constitutionally not yet recognised as a regional or one of the official languages in the Republic of South Africa. However, Northern Ndebele is not going to form part of the discussion in this paper.

The main objective of this paper, is to investigate and discuss the linguistic relationship between isiNdebele of the Republic of South Africa and Zimbabwe. The paper will focus on the (a) historical background and the theoretical views on the origin of the name ‘amaNdebele’ (b) the sound system (c) phonological (d) morphological (e) syntactical and (f) lexical features of the two languages. In conclusion the paper will illustrate how isiNdebele as a language provide the linguistic evidence to some historical facts.

Key Words: AmaNdebele, cross border, Linguistic, morphological, phonological.

Introduction

Historical Background

The Ndebele people of South Africa comprise two main groups known as the Southern Ndebele and the Northern Ndebele. The two groups should not be confused with any other ethnic groups bearing the same name Ndebele, i.e. the Ndebele of Zimbabwe. The members of the latter group are the descendants of Mzilikazi. They have a different history from that of the Ndebele people of South Africa. On one hand, scholars such as Van Warmelo (1930:9), Coetzee (1980:297) and Van Vuuren (1983:13) are of the view that the Ndebele people of South Africa are genealogically related by being the descendants of the same ancestral chief known as Musi. On the other hand, scholars such as Ziervogel (1969) and Jackson (1969: i-iv) have a different views on this matter. Jackson refers to the Northern Ndebele people as the so-called 'Black Ndebele' and he asserts that they do not derive their origin from Musi's tribe. They are the people of Langa, who trace their origin to the erstwhile Zululand just as their Southern Ndebele counterparts do. They are the descendants of a different ancestral chief, Langalibalele. Ziervogel's opinion is in contrast with the latter view; he maintains that these Ndebele people originated from Rhodesia, currently known as Zimbabwe and not from Zululand as is the case with what many scholars have presumed.

The Zimbabwe Ndebeles are a Bantu people found mainly in the western parts of Zimbabwe. Their language belongs to the Nguni-subgroup of the Bantu language family. Their history dates back in the 1820s when they broke away from the then mighty Zulu kingdom, Khumalo (2004:106). Around the 1838-1840, the Zimbabwe Ndebele people came to settle in the southern part of present day Zimbabwe at *Ntabazinduna* near Bulawayo. Historically, the word Ndebele did not therefore refer to a single tribe but to multi-ethnic nations. The Zimbabwe Ndebele language is not a dialect of isiZulu but both are sister languages with a common ancestor which is proto-Nguni.

South African Ndebele is one of the official languages in South Africa while Zimbabwe Ndebele is one of the official languages in Zimbabwe.

Theoretical views on the origin of the name 'amaNdebele'

Taljaard (1993:227) highlights the fact that the Ndebele speaking community refers to the group as 'AmaNdebele' and they proudly refer to their language as '*isiKhethu*' (our language/That which is ours), while Van Warmelo (1930:24) had earlier argued that the amaNala, who speak

one of the Sotho Ndebele dialects, call their language 'isiNdebele', 'isiKhethu', 'isiNdu' (the language of the people) or 'isiNala' (our language). The various etymologies, such as Scholtz's derivation (1957:17), align with what is said above about derivation and suggest that the choice of the name 'Ndebele' does not seem convincing in terms of its origins and history.

Van Warmelo queries the naming of this language as isiNdebele alone, because it has not (as already stated) been successfully explained. However, the origin thereof has been discussed by various anthropologists, historians and linguists, such as Fourie (1921), Bryant (1929), Van Warmelo (1930), Potgieter (1945), Coetzee (1980), Skhosana (1998) and Skhosana (2009). Fourie (1921:26) conducted research on the 'AmaNdebele of Fene Mahlangu' and refers to the word 'Ndebele' as identical to 'Matebele' or 'Amandabili'. He holds three different views in as far as the etymological definition of the concept 'Ndebele' is concerned:

Ilibele en het meervoudig *amabele* beteken borst. Maar met een kleine verheffing van stem in die uitspraak van die laaste lettergreep, beteken hetzelfde woord: kafferkoren. Het werkwoord – *anda* beteken intransitief: toenemen; en transitief: vermenigvuldigen of verspreiden. De tweede *a* van het verbum kan in perfekt-vorm een *e* zijn. Zodat het mogelijk is, dat die naam beteken: zij die de *amabele* verspreid hebben'.

[*Ilibele* and the plural *amabele* mean breast. But if one raises one's voice a little when pronouncing the last syllable, the very same word means sorghum. The verb *-anda* [as an intransitive verb] means to increase and [as a transitive verb it means] to be increased or [to] be spread. The second –*a* of this verb can be an –*e* in the perfect tense, *-ande*. Fourie's first view is that the concept 'Ndebele' means 'people with long breasts' and his second view defines 'Ndebele' as 'people who scatter sorghum'.

The first view is supported by the following praise phrase or accolade of Thabethe (Fourie 1921:25) that says, '*...wa pelela u ma wethu, umfazi o mabele e made*' literally meaning '...our mother is well dressed, the woman with long breasts'.

Fourie's third view is related to a historical approach that is followed by the Nguni group when they name a person after his/her initial founder or ancestral chief, whose name was Ndebele. Thus, the Ndebele people were named after him. Taljaard (1993:227) endorses Fourie's argument, because he regards the Ndebele speaking community as being supposedly derived from 'Ndebele', the man whom the Ndebeles regarded as their great-great ancestor and the founder of the Ndebele tribe.

Bryant (1929:456) and Van Warmelo (1930:7) conducted research on the Ndebele people and advocated that they be known as the Ndebele of the Transvaal who were classified under the sub-group of the Natal Nguni group. These researchers assumed that the concept 'Ndebele' was a Sotho word that the Sotho people used to designate people of Nguni origin. According to Brown (1973:297) in his *Bilingual Setswana-English Dictionary*, however, the verb 'tebela' may also mean 'to strike or knock about with a fist'.

Coetzee (1980:205-7) develops this idea by stating that the 'Ndebele' people were the nameless regiments of Mzilikazi. When they arrived in the Transvaal they came into contact with the Sotho people and were referred to as 'Matebele'. He also regards the name 'amaNdebele' as the equivalent of the Sotho name 'Matebele':

Magubane (2005:8) argues that evidence for the Nguni origins of the Ndebele people is provided by the fact that they are still occasionally referred to as *abaNdungwa* or *baThokwa*, the latter being the Sotho form of this Ndebele term. Both are equivalent to the Zulu term, *abaNtungwa*, which in the 19th century was used to denote members of chiefdoms living in the heartland of the Zulu kingdom. The Southern Ndebele language (isiNdebele or Nrebele in Southern Ndebele) is an African language belonging to the Sotho-Tswana group of Bantu languages, and is spoken by the amaNdebele (the Ndebele people of South Africa). According to Wilkes (2001:312), both the southern and northern groups regard themselves as Ndebele with the former using the name amaNdebele and the latter the name maNdebele when referring to themselves. Magubane (2005:8) attests to the above when he states that the language spoken by contemporary Ndebele-speaking communities provides further testimony to their Nguni origins. Long established patterns of intermarriage between them and their Sotho-speaking neighbours have encouraged the development of very distinctive, hybrid speech patterns that are no longer fully comprehensible to Zulu speakers. Skhosana (2009:18) and Skhosana (2010:137) add that, 'The name Ndebele is commonly used to refer to two genealogically distinct Ndebele groups, namely, the so-called Zimbabwean Ndebele who were the followers of Mzilikazi and are found in Zimbabwe, as their name indicates, and the so-called Transvaal Ndebele who reside within the borders of the Republic of South Africa, comprising two main groups known as the Southern and Northern Ndebele.

The paper investigates the linguistic relationship between the South African Ndebele and the Zimbabwe Ndebele. The investigation will mainly be executed on a synchronic basis. This method of investigation will attempt to expose the linguistic relationship between the South

African Ndebele and the Zimbabwe Ndebele. As the objective is to bring more clarity to the history of the South African Ndebele and the Zimbabwe Ndebele.

Morphological differences between Southern and Zimbabwe Ndebele ‘Noun class system’

An exploration of noun classes serves to illustrate the distinctiveness of the Ndebele groups. Regarding the noun class system, Meinhof and van Warmelo (1932:40) identify 21 noun classes in Ur-Bantu. In most of the Southern Eastern Bantu languages, such as the Nguni languages, some of the noun classes suggested by Meinhof do not occur while others have been reduced to a single noun class. Poulos (1985:16) refers to this process as noun class reduction. Metamorphosis is, of course, inevitable in any living language. Ur-Bantu noun classes that do not occur in the Nguni languages, Southern Ndebele included are Classes **12, 13, 18, 19** and **20** while the content of Classes 16 in these languages has been channelled to Class **17**. In Zimbabwe Ndebele unlike in Southern Ndebele, class **11** and **18** are still in existence. In Southern Ndebele all nouns that were in Class 11 have overtime been channelled to Class **5**. Compare the following examples in this regard:-

Southern Ndebele	IsiZulu	Zimbabwe Ndebele
Class 5	Class 11	Class11
iphawu		uphawu
		uphawu ‘mark’
iphiko	iphiko	uphiko
‘nothing’		
iphondo	uphondo	uphondo ‘horn’
iliZwe	izwe	izwe ‘wood’
Cl 18		Cl 18
-		muva

The variant forms of the Southern Ndebele noun class prefix of Class 5 differ from the Zimbabwe Ndebele in that whilst Southern Ndebele has three variant forms for Class 5, i.e. **i(li)/ il-** and **ilu**, Zimbabwe Ndebele has only two for this class, i.e. **ili** and **-/i-**. In Southern Ndebele, the noun class prefix of Class 5 before polysyllabic nominal stems is: **i-** before monosyllabic **ili** or **ilu** and before vowel verb stems **il-**, respectively, while in Zimbabwe Ndebele the form of this prefix is **li-** in all instances except before stems that commence on the consonant **l-** or on the vowels **i**, **e** or **u**. IsiZulu examples are similar to the Zimbabwe Ndebele ones. Compare the following examples in this regard:

Southern Ndebele

Zimbabwe Ndebele

Cl (5) **a:** iqanda ‘egg’ ithosi ‘drop’

izinyo ‘tooth’

ivila ‘lazy person’

Cl (5) **b:** iliZwe ‘land’

iliva

ilihlo ‘eye’

Cl (5) **c:** ilutjha ‘youth’

ulutho ‘something’

iluju ‘honey’

ulwazi ‘knowledge’

ilusu ‘type of meat’

The form **–lu-** in the Southern Ndebele examples in **(c)** is a relic attesting to the erstwhile existence of Class II in Southern Ndebele. It is important to note that the Class 5 nouns, like those given in examples **(c)** above, may be used with either the prefix **ilu-** or with the prefix **ili-** in Southern Ndebele. In other words a speaker can, for example, either say **iliju** or **uluju**, **ilitjha** or **ilutjha**, **ilisu** or **ilusu**.

Variant form of the Noun class prefixes

In the Southern Ndebele and the Zimbabwe Ndebele as in most other Nguni languages such as isiZulu, isiXhosa, and Siswati, the noun class prefixes comprise two formatives. These formatives are the so-called pre-prefixes, i.e. the initial vowel of the class prefix and the basic or real class prefix that makes the remainder of the class prefix and that which follows on the pre-prefix. Compare the following examples in this regard:

Southern Ndebele

Cl (1) um(u): umuntu

Cl (3) um(u): umuzi

Cl (5) i(li): ithosi

Cl (7) is(i): isitha

Zimbabwe Ndebele

Cl (1) um(u) : umuntu ‘person’

Cl (3) um(u) : umuzi ‘house’

Cl (5) i(li) : ithonsi ‘drop’

Cl (7) is(i) : isitha ‘enemy’

A remarkable difference between Southern Ndebele and Zimbabwe Ndebele as far as the system of noun class prefix is concerned is that the above noun classes have the same variants as is evident in the examples just cited. However, in Classes 8 and 10 Southern Ndebele has the same prefixes and this is a unique feature that is found in Southern Ndebele only. Compare the following examples below:

<p>Southern Ndebele</p> <p>Cl (8): izitja 'dishes'</p> <p>iintulo 'chairs'</p> <p>iimphongo 'foreheads'</p> <p>iiyeleliso 'advice'</p> <p>izono 'sins'</p> <p>Cl (10): iiNkomo 'cows'</p> <p>iimbuzi 'goats'</p> <p>iziNdlu 'houses'</p> <p>iinsimbi 'steels'</p> <p>izimvu 'sheep'</p>	<p>Zimbabwe Ndebele</p> <p>Cl (8): izihlalo 'chairs'</p> <p>izandla 'hands'</p> <p>Cl (10): izinja 'dogs'</p> <p>iziNdlu 'houses'</p>
---	---

What is notable in the examples above, is that in Southern Ndebele Class 8 and 10 has the same prefixes **izi-**, **iin-**, **iim-**, **ii-** and **iz-** as mentioned above. Zimbabwe Ndebele Class 8 and Class 10 have only two class prefixes which are the same but Class 8 prefixes are correctly placed because they emanate from Class 7 nouns, e.g. **isihlalo** (7) > **izihlalo** (8); **inja** (9) > **izinja** (10) respectively.

Unlike other Nguni languages and Zimbabwe Ndebele, Southern Ndebele it has nouns that have the nasal prefixes **iin-**, **ii-** and **iim-** in Class 8. This is another unique feature found in Southern Ndebele. Compare the following examples below:

<p>Cl (7/8): isikhova > iinkhova 'owls'</p> <p>isiphongo > iimphongo 'foreheads'</p> <p>isiwuruwuru > iiwuruwuru 'storm'</p>

In Zimbabwe Ndebele, Class 14 has three variants whereas in Southern Ndebele has only two variants. Compare the following examples below:-

<p>Southern Ndebele</p> <p>Cl (14): ubuso 'face'</p> <p>uboya 'animal fur'</p> <p>utshani 'grass'</p> <p>uboya 'animal fur'</p>	<p>Zimbabwe Ndebele</p> <p>Cl (14): ubukhosi 'kingship'</p>
---	--

Phonological differences between Southern and Zimbabwe Ndebele Mashiyane (2002:79) rightly argues that it is not safe to assume that just because the two nations share the same name there will be no phonetic differences. The following are phonemic

sounds that occur in Zimbabwe Ndebele but not in Southern Ndebele. Compare the examples in this regard:

Southern Ndebele	Zimbabwe
utit tj here	utit sh ala 'teacher'
um tj humayeli	um sh umayeli 'preacher'
tj hayela	tsh ayela 'drive'
tj hada	tsh ada 'get married'

What is notable in the examples above, is that in Zimbabwe Ndebele the voiceless aspirated pre-palatal affricative /**tjh**/ is represented by the alveolar affricative sound /**tsh**/ in Zimbabwe Ndebele.

Southern Ndebele uses the sound /**h**/ in very few words; however, in Zimbabwe Ndebele this sound is generally used. Southern Ndebele mostly uses the voiceless fricative sound **rh**[x]. This is a Sotho sound that was probably borrowed by Southern Ndebele. In Zimbabwe Ndebele this sound is rendered as /**h**/.

Compare the following examples in this regard:

Southern Ndebele	Zimbabwe Ndebele
- rh udula	- h udula 'to drag'
- rh awukela	- h awukela 'sympathise with'
- rh weba	- h weba 'trade'
- rh onona	- h onona 'suspect'
- rh ola	- h ola 'to be paid'

Southern Ndebele uses the aspirated velar explosive sound /**kh**/, while Zimbabwe Ndebele uses the voiced glottal fricative sound /**h**/ in the derivational verbs, i.e. the verbal extensions that emanate from the verb stem **-hamba**. Compare the following examples:-

Southern Ndebele	Zimbabwe Ndebele
- kh amba	- h amba 'go'
- kh ambisa	- h ambisa 'move to another place'
- kh amela	- h ambela 'attend something'
- kh ambelana	- h ambelana 'check other people'

Another common phonological process in Southern Ndebele is denasalization. This denasalization basically implies the omission of the nasal consonant in nasal compounds. According to Meinhof (1932:33), denasalization occurs in many Bantu languages and is usually found in instances where the following sound is retained or may become devoiced, in which case the nasal loses its voicing and eventually falls away. This does not occur in Zimbabwe Ndebele, as can be seen in the examples given.

Southern Ndebele	Zimbabwe Ndebele
ikomo	
inkomo 'beast'	
ikanyezi	
inkanyezi 'star'	
ikalakala	
inkalankala 'crab'	
ikolo	
inkolo 'belief'	
ikunzi	
inkunzi 'bull'	

When one compares the Southern Ndebele and the Zimbabwe Ndebele above, one observes that the same examples given in Zimbabwe Ndebele are also found in other Nguni languages like isiZulu where the nasal is retained in the class prefixes.

Vocabulary of Southern Ndebele vs Zimbabwe Ndebele

While both Ndebele languages have some vocabulary drawn from the isiZulu, research has shown that Zimbabwe Ndebele has adopted more vocabulary than Southern Ndebele as is evident with the nominal stems and the verbal stems below.

Southern Ndebele	Zimbabwe Ndebele	IsiZulu
Nominal Stems		
ipumulo		impumulo
	impumulo 'noise'	
iinhluthu		
izinwele		
	izinwele 'hair'	
ikaba		inkaba
	inkaba 'navel'	
ihloko		
ikhanda		
	ikhanda 'head'	

umkhono			ingalo
	ingalo 'arm'		
imino		iminwe	
		iminwe 'fingers'	
umsana			
umfana	umfana 'boy'		
umseme			icansi
icansi 'grass mat'			
indololwani		indololwane	
		indololwane 'elbow'	
Verbal Stems			
-duda			-
bhukuda		-bhukuda 'swim	
-lotjhisa			
	-bingelela		
	-bingelela 'greet'		
-phunga			
	-chela		
	-chela 'sprinkle with water'		
-gandelela			-cindezela
			- cindezela
'opress'			
-silingeka	-casuka		
		-casuka 'be annoyed'	
-tjhidisa			
	-dedisa		
	-dedisa 'shift something'-		
-kata			-
dlwengula			-dlwengula
'rape'			
-tjhada	-gana		-
gana 'to be married'			

Conclusion

The investigation has shown that although the Southern Ndebele and the Zimbabwe Ndebele people genealogically share the same historical background, origin and the name 'Ndebele', their linguistic status implies that they have developed in two distinct language groups, and are not variant forms of the same language. This paper has conclusively demonstrated that the linguistic dissimilarities that these two Ndebele language (Southern Ndebele and Zimbabwe Ndebele) exhibit, run through all linguistic aspects such as morphology and phonology as well as throughout the vocabulary as shown in examples given prior the conclusion. The differences warrant that Southern Ndebele and Zimbabwe Ndebele are regarded as autonomous languages

that need to be studied or viewed independently. Reviewing the differences in the vocabulary of the Southern Ndebele and the Zimbabwe Ndebele, the conclusion can be drawn that there is insubstantial intelligibility between them. The overall findings revealed that Zimbabwe Ndebele is not a dialect of isiZulu but that both are sister languages with a common ancestry, that is proto-Nguni.

References

- Bryant, A.T. 1929. *Olden times in Zululand and Natal*. London: Green.
- Brown, J.T. 1973. *Setswana English Dictionary*. Gaborone: Botswana Book Centre.
- Coetzee, C.J. 1980. Die strewe tot etniese konsolidasie en nasionale selfverwesening by die Ndebele van Transvaal. Unpublished D. Phil thesis, Potchefstroom: Potchefstroom University.
- Fourie, H.C.M. 1921. AmaNdebele van Fene Mahlangu en hul religieus'- sociaal lewe. D. Phil-proefskrif, Rijksuniversiteit, Utrecht, La Riviere en Voorhoeve, Zwolle.
- Jackson, A.O. 1969. The Ndebele of Langa. Ethological Publication, Department of Cooperation and Development. Pretoria.
- Magubane, P. 2005. *AmaNdebele*. Cape Town: Sunbird Publishers.
- Mashiyane, Z.J. 2006. Beadwork cultural and linguistic significance among the South African Ndebele people. Unpublished DPhil thesis, KwaDlangezwa: University of Zululand.
- Meinhof, C and Van Warmelo, N.J. 1932. *Introduction to the phonology of the Bantu languages*, Dietrich Heimer. Berlin.
- Potgieter, E.F. 1945. Inleiding tot die Klank-en vormleer van IsiNdzundza: 'n Dialek van Suid-Traansvale Ngoenie-Ndebele, soos gepraat in die distrikte van Rayton en Pretoria. Unpublished M.A. dissertation, Pretoria: University of Pretoria.
- Poulos, G. 1985. Typological Trends in South-Eastern Bantu. *South African Journal of African Languages*, 5.(1): 17-23.
- Scholtz, H.v.d.M. 1957. *Taal en Taalverskynsels*. Kaapstad: Nasou.
- Skhosana, P.B. 1998. Foreign interferences in sound, grammatical and lexical system of Southern Ndebele. Unpublished M.A. dissertation, Pretoria: University of Pretoria.
- Skhosana, P.B. 2009. The linguistic relationship between Southern and Northern Ndebele. Unpublished D Litt. Thesis, Pretoria: University of Pretoria.
- Taljaard, P.C. 1993. The history and literature of South Ndebele. In *Comparative Literature and African Literatures*, A.S. Gérard and C.F. Swanepoel, eds., 227-231. Pretoria: Via Afrika.
- Van Vuuren, C.J. 1992. Die aard en betekenis van 'n eie etnisiteit onder die Suid Ndebele, Unpublished D.Phil thesis, University of Pretoria, Pretoria.
- Van Warmelo, N.J. 1930. *Transvaal Ndebele texts*. *Ethnological Publication* 1(2). Pretoria: Government Printer.
- Wilkes, A. 2001. Northern and Southern Ndebele: 'Why Harmonization will not work'. *South African Journal of African Languages*, 21 (3):310-322.
- Ziervogel, D. 1969. Veertig jaar van taalnavoring in Suid- Afrika. *Ethnological and Linguistic Studies in Honour of N.J van Warmelo*. Pretoria: Government Printers.

TEXT DATA INDEXING IN LEXICOGRAPHIC STUDIES: AN APACHE SOLR APPLICATION

B. Tahir Tahirođlu
Çukurova University
Department of Turkish Language and Literature

Abstract

Software technologies developed in recent years have changed the method requirements in linguistic studies and expanded their application areas. Lexicography, which is an applied field of linguistics, also benefits from software technologies. The diversification of online resources that provide data to lexicography requires some form of electronic data saving and indexing of increasing text data. Besides traditional database structures, indexing solutions based on a special data structure model called NoSql have become popular in recent years. It is essential to determine the properties to be questioned in the corpora that will be used for the dictionary to be prepared and to index the data according to this specified template. In this way, presenting the grammatical features to be questioned in a fast and reliable result is one that increases the reliability of the dictionary for the dictionary user and the researcher.

Apache Solr is an open source platform. It is one of the most frequently used platforms for indexing online text data. Compared to traditional database models, the fast operation and relatively high capacity of indexing are among the preferred factors. In the template-based data model, it hosts many modules ranging from tokenization to n-gram extraction and lemmatization processes and allows the user to edit and develop modules.

In this study, the indexing infrastructure used in a neologism project supported by TUBITAK will be given in detail and a discussion will be made about the findings.

Key Words: Lexicography, Corpus Linguistics, Text Data Indexing, Apache Solr, NoSql.

1. Introduction

Lexicography can be defined as a field of research and development that includes the period from the preparation of dictionaries to the presentation and even the pre-lexicography process (Atkins and Rundell, 2008).

Since the 1990s, it can be seen that corpus linguistics, and especially natural language processing (NLP), has expanded its research areas with developing hardware and software technologies. Software technologies have a significant effect in the increase in the number of corpus linguistic resources. Hunston (2002) states that corpus and corpus studies have a revolutionary effect on language studies. Although the aim of NLP studies that are involved in the development of software technologies contributing to this effect is related to the performance levels of the application area, problem-solving and problem-solving in an algorithmic plane form both the sub-branches of NLP and the common working area of NLP.

Hartmann and James (1998) described the corpus as a systematic collection of texts that were brought together to examine language characteristics and language variations. Today, creating a corpus and analyzing all kinds of features from the corpus has become two separate research topics. Creating a corpus can also be seen as a statistical problem. It is also the subject of statistical sampling theory that the representation and balance of a written language corpus can be established. In this respect, when the scaling and measurement researches of other social sciences are considered, linguistic studies have been ignored, but lately, the language features have been dealt with in a statistical and computational way. Considering vocabulary, each word or even each character is in a statistical behavior, and as the number of word entities that make up the collection increases, the word embedding method becomes more evident in the machine learning, where these behaviors show a wide range of distant or close dependencies from patterns to models.

Dictionary preparation, dictionary based on data; it is important to define the units in the dictionary based on the data and to obtain the characteristics of the words from the big data in terms of the reliability and comprehensiveness of the dictionary. If the dictionary to be prepared covers all areas of the general dictionary, the data on which the dictionary is based should be collected from various sources. Atkins and Rundell (2008) describe dictionaries as a kind of database. It can be said that every dictionary can be considered in such a way as computational thinking or computational thinking. Thus, the whole structure of the dictionary is handled step by step in an algorithm pattern so that the produced result is either printed format or a dynamic web page or both.

Although today's database structures have the capacity to carry all the representations of a dictionary to be prepared, besides the classical SQL-based database structures, different document-based software of the classical database architectures called NoSQL have been released. These software provide advantages, especially in the process of saving, indexing and querying text data.

In this study, it will be mentioned how Apache Solr indexing and querying system used in the scope of a project¹ supported by TUBITAK has been customized during the creation of project word lists and extraction of the contextual index.

2. Apache Solr Indexing and Querying System

Indexing in a collection is a process that allows words, phrases, or labels to be easily accessed without re-scanning the entire corpus (McEnergy & Hardie, 2012). The text data on the Internet must be indexed in order to be converted into searchable form by search engines. The search engines record all units, especially words, in separate lists on the web pages they index, and contain details such as the page information of each list and the unit. In this way, the search engine's web crawler software can detect the creation time of web pages. They can also calculate changes in previously indexed pages by calculating their properties. In short, indexing saves both time and storage space for saving data. It is a technology that facilitates access to a whole page or resource from words.

¹ New words (Neologism) in Turkish of Turkey's only online News Text Automatic Extraction, TUBITAK-SOBAG 111K223 numbered project. We thank TÜBİTAK for their support.

Apache Solr is an open source search and indexing platform based on the Apache Lucene infrastructure. 4.6 version of the software is used in our project. The platform has now reached version 8.1.1, and naturally new features have been added to the platform.

3. Using Apache Solr for the New Words Project

Turkey Turkish was used in the online news text Turkish Language Society to take place in the current Turkish dictionary new words used for automatic extraction Apache Solr granting the candidate of new words in the context obtained and word frequency. The aim of the project is to determine the new words from the data obtained from the crawling and indexing of all pages of Turkish online news sites to a depth of 1000 links for 3 years. For this purpose, more than 600 different web sites have been identified and these sites have been scanned with Apache Nutch web crawler, which is Apache open source software. Compliance with Nutch and Solr made it much easier for the project to collect data. The data obtained at the end of the project exceeded one terabyte.

Solr is preferred because it is open source and can deliver data quickly in large-scale projects and allows distributed architecture. The data collected were divided into eight shards in our project. Solr was accessed via the web browser and operations related to the data were performed on the address line. Having a web browser allows using Solr without the need for any other server software.

The following screenshot shows the features of the Solar used in the project. The total number of documents indexed is 14.436.408, ie the number of web pages scanned by this project.

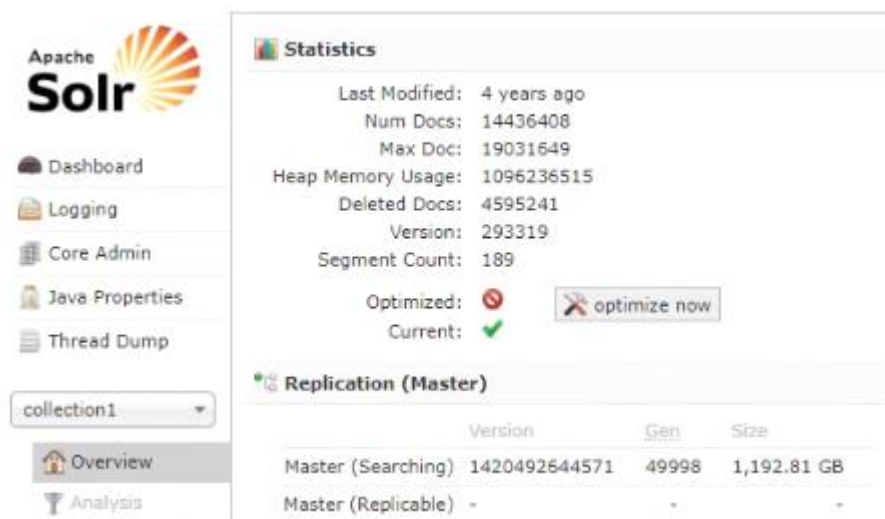


Figure 1. New Words Project Apache Solr Software Overview Screen

It is known that modern lexicography research is indispensable for specialized software in both data collection and processing stages. In this respect, Apache Solr also has NLP tools that a lexicographer can use to index large text data. Tokenization, stemmer, n-gram extraction, term frequency extraction, predetermined classification structure at the time of parsing and the parsed section allows the opportunity to record and query them. Solr, which can be used by the lexicographer as a basic database, can also provide input to any application from the address line.

As an XML schema structured platform, it is possible to enrich the schema with new elements according to the needs by stretching the existing schema structure by default. In our project, verbs, adjectives and adverbs fields are added to the existing schema to determine how much the existing dictionary lists are used in the indexed data. Solr is a software that executes operations according to the item in the XML file. At the time of indexing, it is determined what to save and calculate according to the preconfigured schema elements. It has the so-called "Factory" features, especially the functions ending in text processing language. For example, the following screenshot shows the structure of the function that extracts 2-n-gram words in the schema file.

```

<!--ngram alani-->

<fieldtype name="ngram" stored="true" indexed="true" class="solr.TextField">
<analyzer type="index">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.ShingleFilterFactory" maxShingleSize="2" outputUnigrams="false"/>

</analyzer>

<analyzer type="query">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.ShingleFilterFactory" maxShingleSize="2" outputUnigrams="false"/>

</analyzer>

<!--ngram alani bitis-->

```

Figure 2. Function Structures for Apache Solr n-gram Extraction

As can be seen in the figure above, class elements offer filters that enable both tokenization and lowercase conversion. The analyzer field is mandatory for each of the specified fields. The word n-gram in Solr is extracted with the solr.ShingleFilterFactory class.

```

<!--isimler_tekli alani baslangic-->
<fieldtype name="isin_tekli" stored="true" indexed="true" class="solr.TextField">
<analyzer type="index">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.KeepWordFilterFactory" words="isimler.txt" ignoreCase="true"/>
</analyzer>

<analyzer type="query">
<tokenizer class="solr.WhitespaceTokenizerFactory"/>
<filter class="solr.LowerCaseFilterFactory"/>
<filter class="solr.KeepWordFilterFactory" words="isimler.txt" ignoreCase="true"/>
</analyzer>

<!--isimler_tekli alani bitis-->

</fieldtype>

```

Figure 3. Nouns Field and Properties

The screenshot above shows the class of functions that extract names from a text. Parsing from spaces is used for tokenization and a list file is given as source. According to the list in the file,

the most frequent nouns were saved according to their frequency and made searchable at any time. The following screenshot also shows the top 10 lists of names by frequency.

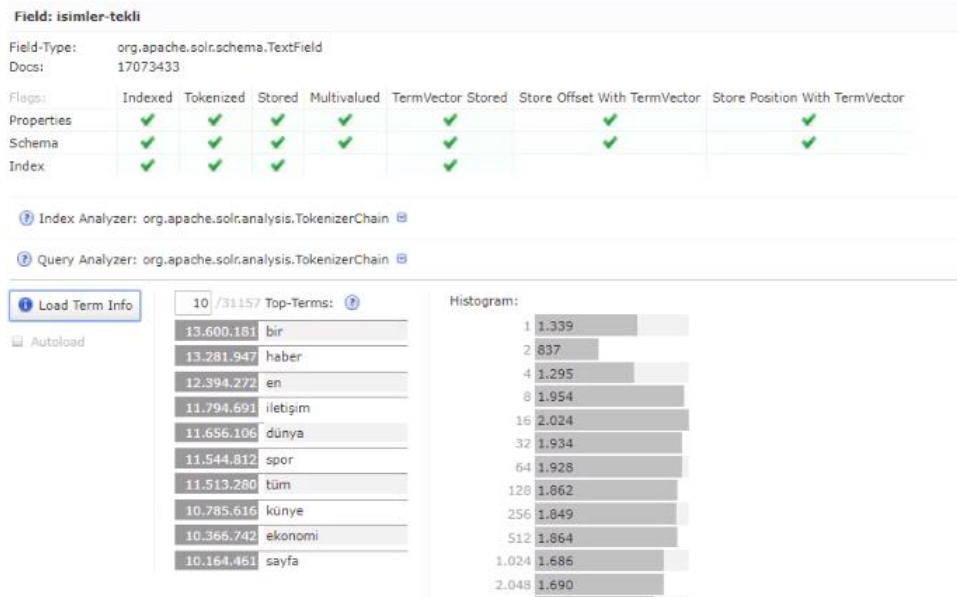


Figure 4. Top 10 “Nouns” Frequency Display in Apache Solr

Within the scope of the project, in addition to the determination of new words, the main purpose of the project was the extraction of the sentences with the regular expression structure in the Turkish news texts in line with the opportunities provided by Solr's text processing tools. These are recorded as additional gains from the project. From this data compiled between 2011 and 2014, it was tried to obtain as detailed lexical segmentation and views as possible.

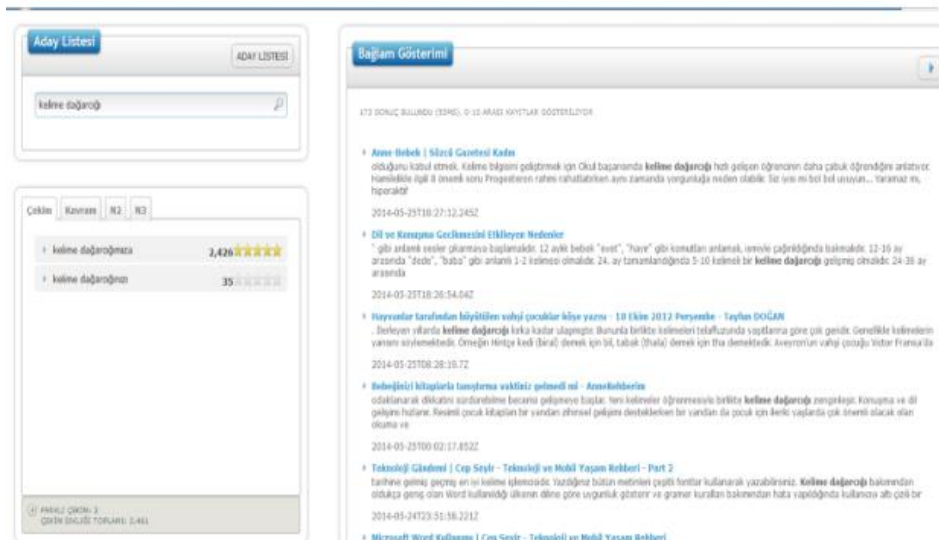


Figure 5. Apache Solr Word Context Display

The results of the queries sent to Solr were used to illustrate the context of the new words obtained from the project data. In the example above, the context representation of “kelime dağarcığı” is obtained from the data recorded in Solr using the highlight (hl) parameter.

```

<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">2108</int>
    <lst name="params">
      <str name="q">açıklık</str>
      <arr name="fl">
        <str>idf(text,açıklık)</str>
        <str>tff(text,açıklık)</str>
      </arr>
      <str name="rows">1</str>
    </lst>
  </lst>
  <result name="response" numFound="92505" start="0">
    <doc>
      <double name="idf(text,açıklık)">6.009498119354248</double>
      <long name="tff(text,açıklık)">144743</long>
    </doc>
  </result>
</response>

```

Figure 6. IDF Score and Frequency of Use of the word “açıklık“

The fast and useful use of the Solr indexing and querying platform in big data simplifies the work of the lexicographer, while the copy content problem of the web continues in Solr. The fact that copy content exists in many sources is seen as a problem that needs to be overcome in the text preprocessing phase. Although the Nutch web crawler scans different URL addresses and saves them to Solr as different URL addresses, the problem of repeating the content saved as the “content” field has not been resolved effectively.

4. Conclusion

The Internet provides us with huge amounts of text data. Electronic environments that open new horizons for lexicography (documents such as web, pdf, etc.) have enabled us to look at text data from different perspectives with new algorithms. Considering that each digitized text consists of computable words and a network of these, it is clear that the lexicologist must rethink the latest problem-solving algorithms for software use and classification.

Internet is an environment where billions of words build online networks every day. To be able to follow and use the new features of the software platforms to be used for the samples to be obtained from this environment, adding new ones to the existing features will both facilitate the preparation of dictionaries and improve the vocabulary experience of the dictionary user.

References

- Atkins, B. T. and Michael Rundell (2008), *The Oxford Guide to Practical Lexicography*, Oxford University Press, New York.
- Hartmann, R.R.K. and Gregory James (1998), *Dictionary of Lexicography*, Routledge, New York.
- Hunston, S. (2002) *Corpora in applied linguistics*, Cambridge University Press, , Cambridge.

McEnery, Tony and Andrew Hardie (2012), *Corpus Linguistics, Method, Theory and Practise*, Cambridge University Press, , Cambridge.

<https://lucene.apache.org/solr/>

<https://nutch.apache.org/>

AN OVERVIEW ON THE HISTORY OF RUSSIAN LEXICOGRAPHY

Lecturer Ümmügülsüm DOHMAN

Trakya University, Faculty of Letters, Department of Translation and Interpretation

Abstract

The dictionaries prepared between the 11-17th centuries are the first period of Russian lexicography and have the names of “*Glossary*”, “*Alfavit*”, “*Azbukovnik*”, “*Lexicon*”. These dictionaries include the translation of an unfamiliar word or words that are difficult to understand, the interpretation of their meanings, and the explanation of concepts, specific names and symbols in any work. The development period of Russian lexicography, which has a rich dictionary tradition, is the 18th century. In this period, the Russian language, culture and literature were on the rise. The most brilliant period of Russian lexicography was the 19th century during which Russian vocabulary was recorded intensively. With the 19th century, the type, volume and functions of the dictionaries changed, there were important developments in the field of lexicography, and basic dictionaries of Russian language were compiled. The 20th century advanced on the foundations that it took from the 19th century more steadily. The success of the 20th century stems from the fact that the first lexicography theories were formed in this period. This research applied a descriptive qualitative approach on the history of Russian lexicography. The main objective of this research is to describe the historical development of Russian lexicography from the beginning to the end of the 20th century. In this context, we will examine lexicographical studies in Russia in light of the structure of the period in general terms under two headings as the 11th – 17th centuries and the 18th, 19th, 20th centuries. Moreover, information will be given about the prominent explanatory dictionaries of each period.

Key Words: Russian lexicography, history of lexicography, dictionaries, theories of lexicography

1. Introduction

The history of Russian lexicography, which has a rich dictionary tradition, is usually divided into several periods by researchers. In this context, O. L. Rubleva divides the developmental stages of Russian lexicography into three periods (Rubleva, 2004, p. 28).

1. The pre-dictionary period: the 11th – 15th centuries
2. The early dictionary period: the 16th – 17th centuries
3. The dictionary development period: the 18th – 19th – 20th centuries

In the study *The History of Russian Lexicography* (История русской лексикографии) prepared by L. S. Kovtun, O. D. Kuznetsova, G. N. Skljarevskaja and F. P. Sorokoletov, Russian lexicography was divided into two periods, and the second period itself was divided into three stages (Kovtun, Kuznetsova, Skljarevskaja and Sorokoletov, 2001).

1. The old period: the 11th – 17th centuries

2. The new period (the period from the 18th century to the 90s of the 20th century)

2.1. The 18th century

2.2. The 19th century

2.3. The 20th century

In the lexicography entry stated in his encyclopedia the *Russian Language* (Русский язык), linguist Yu. N. Karaulov (1935-2016) examined the Russian lexicography in five terms in detail (Karaulov, 1997, p. 209-211).

1. The first period: the 11th -15th centuries

2. The second period: the 16th -17th centuries

3. The 18th century

4. The 19th century

5. The 20th century

When we look at the developmental periods of Russian lexicography mentioned above, it can be seen that the history of Russian lexicography began in the 11th century and 18th century was expressed as the development period. The fact that the confusion about language in Russia was still continued in the 17th century, the main lines of the Russian literary language began to form gradually (Ozer, 2004, p. 103) and the spelling reform took place at the I. Peter period in the 18th century had a very significant effect on this. After the spelling reform, the public adopting this language reform, applying it and determining the rules of the Russian language was not naturally straight away. Therefore, Russian lexicography began to develop slowly in the 18th century, had its liveliest period in the 19th century, and continued to rise in the following years. We will also examine lexicographical studies in Russia under two headings as the 11th – 17th centuries and the 18th, 19th, 20th centuries.

2. Methodology

This research applied a descriptive qualitative approach on the history of Russian lexicography. In this context, in order to create a database to be used in the current study, the field of Russian lexicography literature was reviewed. Both printed and non-printed books, academic journals and electronic databases were reviewed for this study.

3. Objective of the research

The main objective of this research is to describe the historical development of Russian lexicography from the beginning to the end of the 20th century. The historical development of Russian lexicography will be examined in light of the structure of the period in general terms, and information will be given about the prominent explanatory dictionaries of each period.

4. Results and Discussion

4.1. The 11th – 17th Century Russian Lexicographical Studies

The first dictionary which belongs to the Eastern Slavs to date is the list of words in the *Novgorod Church Rules Book* (1282) (Новгородская Кормчая). The dictionary which is composed of 174 words, includes the words of the Hebrew and Greek words encountered in the holy books, some words from other languages, and even old Slavic words. Thus, it is clear that the first Russian

dictionary is composed of words which are from religious texts but are unclear (Dubichinskiy, 2008, p. 18). The second famous dictionary is *Novgorod Dictionary* (Новгородский словарь) which was prepared in 1431. Russia's acceptance of Christianity in the 10th century influenced Russian vocabulary and caused the words of Greek, Hebrew, Latin and especially the Church language to pass into the Russian language. Therefore, as mentioned earlier, the first dictionaries include the unclear words from religious texts.

At the end of the 16th and 17th century came to exist dictionaries are called as the type of "Azbukovnik". With the emergence of azbukovniks, the word lists previously given depending on the text are now given in alphabetical arrangement. For this reason, dictionaries are called "Azbukovnik" meaning alphabet. The first printed "Azbukovnik" was written by the famous cleric, linguist, writer and translator Lavrentiy Zizaniy Tustanovskiy (1550-1634) in Vilnius in 1596 and it was called *Lexis* (Лексис). In this dictionary which contains 1061 words, the explanations include the words of other languages, especially the Church Slavic words (Fomina, 1990, p. 17). Another important dictionary of the 16th century was prepared by the linguist and writer Maksimus the Greek (1470-1556). This work is called *Psalter* (Толковая Псалтырь) and it is known that it was translated from Greek into Russian by M. the Greek and his colleagues Dmitry Gerasimov and Vlas Ignatov from Novgorod (Kozyrev and Chernjak, 2000 p. 16). Dubichinskiy, working on Ukrainian and Russian lexicography, states that the cultural developments that affected Moscow Russia at that time, the new lexicography principles and the rules for the creation of a dictionary came from South-West Russia (Dubichinskiy, 2008, p. 20). In other words, these dictionaries, which are mentioned in Russian lexicography as the first Russian dictionaries, originated from Kiev, and the fact that the lexicographers are mostly Ukrainian and the dictionaries are printed in cities such as Kiev and Vilnius prove this view.

From the second half of the 16th century to the 17th century, interest in truth and concrete matter increased, and the desire to reach real knowledge increased the curiosity for encyclopedias. Thus, azbukovniks emerged as encyclopedic dictionaries. Descriptions like person names, place names, basic meaning of words, country, folk, plant, animal etc. names are described in azbukovniks. The fact that the azbukovniks contain a variety of information became the basis for the preparation of different types of dictionaries. Thus, the 17th century it is a period when the foundations of the dictionary of the century were laid.

Considering the dictionaries created in the 17th century, it can be noticed that the number of bilingual and multilingual dictionaries increased. For example, the dictionary (Лексикон славеноросский и имен толкование) was prepared in 1627 by the famous Ukrainian linguist, author, translator, printer and pedagogue Pamva Berynda (1555-1632), and Greek and Latin handwritten lexicons were prepared by teacher, translator, lexicographer and cleric Epifaniy Slavinetskiy (1600-1675) (Kovtun, et al., 2001, p. 55). The dictionaries prepared between the 11-17th centuries are the first period of Russian lexicography and have the names of "Glossary", "Alfavit", "Azbukovnik", "Lexicon" as understood from the dictionary titles given above. These dictionaries include the translation of an unfamiliar word or words that are difficult to understand, the interpretation of their meanings, and the explanation of concepts, specific names and symbols in any work.

4.2. The 18th, 19th, 20th Century Russian Lexicographical Studies

The 18th century was a new era for Russia. In the first quarter of the 18th century, the spelling reform was implemented by the order of Peter I. Peter I was known to give importance to the

preparation of the dictionary. Thus, the first foreign words dictionary was prepared with the order of I. Peter himself and it was called *Dictionary of New Words in Alphabetical Order* (Лексикон вокабулам новым по алфавиту). This dictionary was the first dictionary in the 18th century in which Russian words were taken from foreign languages. The dictionary, which was supposed to be prepared in 1725 or before 1725, is incomplete and unprinted (Birjakova, 2010, p. 29, 30).

Another important dictionary of the 18th century was the *Dictionary of the Russian Academy* (Словарь академии российской) (1789-1794) published in six volumes between 1789 and 1794. Yekaterina Romanovna Dashkova (1743-1810), who was interested in different fields such as philosophy, linguistics and lexicography and has prepared an establishment project of the Russian Academy, took an active role in the creation of this dictionary. The Russian Academy which opened in 1783, began its work with the idea of creating a dictionary that correctly defines the Russian language and aims to put the Russian words in the foreground instead of foreign words. This study is considered as not only the first explanatory dictionary of the Russian language but also the first Academy dictionary.

In the 19th century, Russia was virtually oblivious to education, linguistics, literature and lexicography. Poet and writer Aleksandr Sergeyeovich Pushkin gave Russian the *Golden Age* with the works he produced in the first quarter of the 19th century. In this period Russian lexicography was marked by the emergence of a number fundamental explanatory dictionaries. Considered as the founder of the Russian literary language, the language used by Pushkin in his works determined the language of contemporary Russian. Thus, the language used by Pushkin formed the vocabulary of the dictionaries prepared in this period. That's why the 19th century was the most brilliant period of the beginning of Russian lexicography. The 19th century dictionary works began with the republishing the *Dictionary of the Russian Academy* (Словарь академии российской) in the years 1806-1822. With the first publishing of this dictionary, the explanation and description of the Russian language gained importance. In addition to this, the first dictionaries of etymology and dialectology types were prepared in the 19th century.

In the second half of the 19th century, doctor V. I. Dahl prepared the explanatory dictionary called *The Explanatory Dictionary of the Great Russian Language* (Толковый словарь живого великорусского языка) (Dahl, 1801-1866). This dictionary, which is the most famous among the dictionaries of dialect, was published in 4 volumes between 1863 and 1866. The first edition of this dictionary included more than 30,000 proverbs, idioms, riddles. On the first page of the dictionary, Vladimir Dahl emphasized that the title of dictionary used the word "explanatory" (толковый) since not only the translations of the entry words but also their explanations were given. The headwords in the dictionary are arranged according to the root word method, and the dictionary includes the Russian spoken in the first half of the century. The dictionary, which was printed many times and kept up to date, continues to be published as a single volume today.

The 20th century advanced on the foundations that it took from the 19th century more steadily. However, the 20th century was a difficult period in which Russia changed in many areas such as the state structure, policy and language. Despite this difficult and complicated period, Russian linguistics and Russian lexicography began to be formed in theory. L. V. Shcherba (1880-1944) with his study *Towards a General Theory of Lexicography* (Опыт общей теории лексикографии) (Shcherba, 2004) in 1940, S. I. Ozhegov with his study *Lexicological and Lexicographical Problems* (Вопросы лексикологии и лексикографии) (Ozhegov, 1973) in 1953, and V. V. Vinogradov with his study *On Some Problems of Russian Dictionary* (О

некоторых вопросах теории русской лексикографии) (Vinogradov, 1977) in 1956 provided the theoretical foundations of Russian lexicography.

As mentioned earlier, 20th century was a complex and difficult period for Russia. Factors such as Russia's participation in the First World War, followed by the October Revolution of 1917 and the rapid change of society and economy in connection with this naturally affected Russian vocabulary. In addition, the Russian language once again experienced the spelling reform in the Soviet era of 1917-1918. All these reasons led to the change of Russian language. It was necessary to prepare a dictionary to record and explain the changes of the era. For this purpose, D. N. Ushakov (1873-1945), who began to work in 1928, published four volumes of the *Explanatory Dictionary of the Russian Language* (Толковый словарь русского языка) between 1935-1940 (Ushakov, 1996). In this dictionary the origin of words was also mentioned. In 1949, S. I. Ozhegov (1900-1964), who actively worked in Ushakov's dictionary, published the work of the *Dictionary of the Russian Language* (Словарь русского языка) (Ozhegov, 1988). The dictionary of Ozhegov's, which contained the contemporary Russian literary language, benefited from Ushakov's dictionary. This study which is the first single-volume dictionary of the Russian language, includes only words used in the 20th century. While in Ushakov's dictionary, the origin of the words is shown, Ozhegov's dictionary does not include the origin of the words. After the death of Ozhegov, the new editions of the dictionary were published by the editor N. J. Shvedova (1916-2009) who was a Russian literary historian and linguist.

Between the years of 1950-1965, the Academy of Sciences published a dictionary with 17 volumes whose first edition was published under the title of *A Dictionary of Contemporary Russian Literary Language* (Словарь современного литературного языка), and second and third editions were published under the title of *Big Academic Dictionary* (Большой академический словарь русского языка). The dictionary covers the period from the 19th century to the 21st century, from Pushkin to the present. It is a large-scale study that reflects the contemporary Russian literary language and Russian vocabulary. The editions of this dictionary are still going on.

From 1957 to 1961, the 4-volume *Russian Dictionary* (Словарь русского языка), also known as *Small Academic Dictionary* (Малый академический словарь), was prepared under the editorship of linguist and lexicographer A. P. Evgenjeva (Evgenjeva, 1985). The study, which aimed to reflect the vocabulary of the literary language, covered the period from Pushkin to the present. 20 years after the publication of this dictionary, the second edition was published in 1981 under the editorship of A. P. Evgenjeva. In this edition, the recording of new words entering the language, the update and addition of examples were enabled within the last 20 years.

Based on the fact that words are variable mechanisms, the Russian Academy of Sciences composed a series of dictionaries called the *Dictionary of New Words and Meanings* (Новые слова и значения. Словарь-справочник) by following new vocabulary beginning from the 1960s and also the vocabulary transferred from other foreign languages into the Russian language. For this study, which is still ongoing, periodicals are tracked at certain time intervals such as periods of 10 years, and the changes in words, new formations and new words transferred into the language in 10 year processes are identified and recorded in dictionaries. For example, new words are studied and recorded in periodicals and literary works published between the years of 1950-1960 and 1960-1970. In this context, the first dictionary of new words recorded in 1971 is titled *New Words and Meanings. Composed According to the Press and Literary Sources of the 60s*.

Dictionary, Reference Book (Новые слова и значения. Словарь-справочник по материалам прессы и литературы 60-х годов) (Butseva, Levašov and Denisenko, 2009). The second of this series covered 70s years and was published in 1984. The third one contained the new words of the '80s, and the fourth included the new words of the '90s and was published in 2009. The dictionary of the new words formed in the first 10 years of the 21st century is also being prepared. Many of these words are included in explanatory dictionaries. These words are sometimes included in the revised editions of dictionaries. Apart from these series, dictionaries on new words and foreign originated words are being prepared under the titles of new words dictionary and foreign words dictionary.

In 1998, *The Big Explanatory Dictionary of the Russian Language* (Большой толковый словарь русского языка) was prepared under the editorship of the linguist and lexicographer S. A. Kuznetsov (Kuznetsov, 2000). The dictionary, which continued the dictionary tradition of the academy, contained approximately 130,000 words. The dictionary containing the words of the 19th century explained the terms in literary works, scientific popular publications, written media, spoken language, modern science and technique, economy, art, philosophy, law and history. However, in the last years of the 20th century, it became popular and included words in the fields of astrology, parapsychology and religion, which were included in the Russian literary language. The origins of the words quoted in Russian from the foreign languages were immediately shown in parentheses after the beginning of the article.

Again in 1998, *The Russian Language Explanatory Dictionary of the End of the XXth c. Language Changes* (Толковый словарь русского языка конца XX века. Языковые изменения) was published under the editorship of G. N. Skljarevskaja (Sklyayerevskaja, 1998). This study aimed to show the developments that occurred in Russian vocabulary between the years 1985 – 1997. The dictionary contained completely new and new words that have gained new meanings. In the preface of the dictionary, it is stated that some of these new words exist in the new editions of Ozhegov's dictionary and in *Big Academic Dictionary*. The origin of the words was not shown in the dictionary.

5. Conclusion

The dictionaries prepared between the 11th and 17th centuries which refer to the first period of Russian lexicography are called “*Glossary*”, “*Alfavit*”, “*Azbukovnik*”, “*Lexicon*” and they are mainly from South-West Russia. The first dictionaries prepared were lists of irregular words that were not understood in religious texts in general.

With the spelling reform in the 18th century and then with the opening of the Russian Academy, a new era began for Russian lexicography studies. However, the face of the lexicographers changed. In the previous periods, the lexicographers were mostly religious men such as Maksimus the Greek, Pamva Berynda and Lavrentiy Zizaniy Tustanovskiy, but with the 18th century, professors, teachers, academicians, translators, government and community men were among the lexicographers. 19th century in Russian lexicography was marked by the emergence of a number fundamental explanatory dictionaries like the dictionary of the Russian Academy and Dahl's dictionary.

Although the 20th century was a period in which Russia experienced innovations and difficulties in many fields such as social, linguistics and politics, the theoretical foundations of Russian lexicography were formed in this period. Thus, dictionaries were prepared according to clearer

principles and methods from then on. With the advancement of technology, it was a period in which the words of English origin were included in the Russian language, therefore the dictionaries were updated, or dictionaries based on new words were created. The 20th century continues to remain in memory and maintain its existence as a brilliant period in Russian lexicography with the existence of dictionaries and lexicographers who are still up-to-date.

References

- Birjakova E. E. (2010). *Russkaja leksikografija XVIII veka*. S.t. Petersburg: Nestor-İstoriya.
- Butseva, T. N., Levašov, E. A., and Denisenko, Yu. F. (Eds.). (2009). *Novije slova i značenija. Slovar'-spravočnik po materialam pressı i literaturu 90-x godov XX. veka v dvuh tomah*. St. Petersburg: Ran.
- Dahl, V. I. (1801-1866). *Tolkovij slovar' zhivogo velikoruskoga jazika*. (4 vols.) Moscow.
- Dubichinskiy, V. V. (2008). *Leksikografija ruskogo jazika*. Moscow: Nauka, Flinta.
- Evgenjeva, A. P. (1985). *Slovar' ruskogo jazika v cetireh toma*. (3rd edn.) Moscow: Russkij jazik.
- Fomina, M. I. (1990). *Sovremennij ruskij jazik. Leksikologija*. Moscow: Viššaja škola.
- Karaulov, J. N. (1997). *Ruskij jazik, entsiklopedija*. Moscow: Drofa.
- Kovtun, S. L., Kuznetsova, O. D., Sklyarevskaya, G. N., and Sorokoletov, F. P. (Eds.). (2001). *Istorija ruskog leksikografii*. St. Petersburg: Nauka.
- Kozyrev V. A., and Chernjak, V. D. (2000). *Vselennaja v alfavitnom porjadke: Očerki o slovarjah ruskogo jazika*. S.t. Petersburg.
- Kuznetsov, S. A. (2000). *Bol'shoj tolkovij slovar' ruskogo jazika*. St. Petersburg: Norint.
- Ozer, Z. B. (2004). *Rus Dilinin Gelişme Evreleri*. Ankara: Çetin Ofset.
- Ozhegov, S. I. (1988). *Tolkovij slovar' ruskogo jazika*. (20th edn.). Moskva: Ruskiy jazik.
- Ozhegov, S. I. (1973). *Leksikologija, Leksikografija, kultura rechi*. Moscow: Viššaja Skola.
- Rubleva, O. L. (2004). *Leksikologija sovremennogo ruskogo jazyka*. Vladivostok.
- Shcherba, L. V. (2004). *Jazikovaja sistema i rechevaja dejatel'nost'*. (2nd edn.). Moscow: Editorial URSS
- Sklyarevskaja, G. N. (1998). *Tolkovij slovar' ruskogo jazika kontsa XX veka. Jazykovie izmenenija*. St. Petersburg: Folio - Press.
- Slovar' Akademii Rossijskoj. 1789-1794*. (6. Vols.) S.t. Petersburg.
- Ushakov, D. N. (1996). *Tolkovij slovar' ruskogo jazika*. (4 Vols.) Moscow: Terra.
- Vinogradov, V. V. (1977). *Izbrannije trudi. Leksikologija i leksikografija*. Moscow: Nauka

CREATING A TRILINGUAL DICTIONARY FOR WESTERN YUGUR, AN ENDANGERED TURKIC LANGUAGE

Yarjis Xueqing Zhong

The Australian National University

Abstract

The Yugur are one of the smallest ethnic minorities in north-western China. They mainly speak two distinct endangered languages, Western Yugur (a Turkic language) and Eastern Yugur (a Mongolic language). This paper discusses at least three challenges in compiling a comprehensive ethnographic Western Yugur-English-Chinese (online) trilingual dictionary, with the main audience being the Western Yugur community, as well as Chinese-speaking and English-speaking researchers on Turkic languages. The first challenge is the importance of creating a practical and usable orthography that is acceptable to the local language community. The International Phonetic Alphabet (IPA), which has been used in a few Yugur language research publications, is not easy to learn or use. Throughout China, people learn to read Mandarin characters and the pinyin romanization. This renders both the IPA and the modern Turkish alphabet less variable as orthographies for Western Yugur. Consequently, the proposed orthography uses a Latin alphabet, mainly based on the pinyin system, which accommodates not only regional variants, but can be extended to Eastern Yugur with some additional symbols. The second challenge involves defining and translating words, arranging words in semantic domains, providing definitions within the Western Yugur language itself, and providing syntactical frames for words. Vernacular definitions in the style of the Collins COBUILD dictionary simultaneously illustrate a syntactic frame of a word and its meaning. The third challenge stems from the importance of having a good online dictionary. Compiling a comprehensive dictionary with sounds and pictures will allow for spin-offs which could be republished in parts for a children's dictionary, teaching materials at schools or other language revitalization purposes. However, there needs to be a lot of time and some funding to develop a good online dictionary, especially a crowdsourcing version, although currently we use Webonary as a free basic draft online dictionary.

Key Words: Western Yugur, endangered language, trilingual dictionary

Creating a Trilingual Dictionary for Western Yugur, an Endangered Turkic Language

The Yugur mostly reside in Sunan Yugur Autonomous County within Gansu Province in the northwest of China. According to the 2010 Census (National Bureau of Statistics of China, 2011), there were 14,378 Yugur people. Yugur people mainly speak two distinct languages: Western Yugur (also known as Saryg Yugur), a Turkic language, and Eastern Yugur (also known as Shera Yugur), a Mongolic language, both with about 2,000 speakers (Y. X. Zhong, 2019); while other Yugur people only speak Mandarin Chinese, which is the lingua franca.

It has been claimed that the most helpful field research product for an endangered language speech community is creating a dictionary (Ogilvie, 2011). Since Western Yugur has a declining number of speakers and its vocabulary has not been comprehensively documented, a well-designed dictionary, especially an online dictionary, is sorely needed to help preserve and maintain the language and culture. Furthermore, the rich complexity of the language itself and the community's desire for language maintenance have required the latest approaches to dictionary making. Therefore, the aim is to create a Western Yugur-English-Chinese dictionary (including an online dictionary) using the methods of modern practical and theoretical lexicography. The collection of material for the dictionary, the structure of the dictionary and the design of the entries were based on a solid grounding in lexicography and dictionary-making for endangered languages, as well as the desire to create a dictionary, which will arouse the curiosity of native speakers to connect with their identity by learning their language, and enhance their motivation to do so. The ultimate aim of creating such a trilingual dictionary is to preserve the richness of the Western Yugur language, allowing it to be maintained and used by communities and passed onto younger generations, and also to make the language accessible to both Mandarin Chinese-speaking and English-speaking learners and researchers.

This paper focuses on discussions of some of the processes and challenges of creating such a dictionary, which includes data collection and selection methods, some of the compilation processes and challenges for an endangered language dictionary project, including the potential for making a crowdsourcing dictionary.

Methods

Some practical considerations and most elementary issues of the dictionary determine the actual use of the dictionary for the target audience, developing a practical orthography if there is not one, and determining the design, presentation and organization of the material. It is here that the most fundamental decisions are made, such as identifying the target audience for the dictionary; formatting of the dictionary and so on, which can affect every subsequent aspect of the dictionary-making (Atkins & Rundell, 2008). Therefore, it is essential to construct the dictionary project with a specific plan in mind before the data collection and selection stage. The discussion of this section is from the perspective of a native speaker of an endangered language, where the approaches and steps include collecting words from fieldwork, grouping words by semantic domains, selecting words and example sentences from a small-scale corpus, and reviewing existing publications, incorporating material from old word lists and checking them with modern speakers.

Fieldwork

Fieldwork data is the most critical resource for not only study of the language itself but also specifically for the entries and example sentences in the dictionary. The primary methods for this paper include consultation, interviews, elicitation, participant observation and recording of narratives/free speech. Before or while collecting the data, the project requires extensive consultation with a broad range of people who are the dictionary's audience, and to incorporate their suggestions for the dictionary format, contents and outcomes. Eventually, a draft outline of the dictionary was created based on the criteria and checked again with some potential users to ensure that it could realistically meet their expectations.

The primary method of data collection is using interviews to elicit meanings of words, including structured and semi-structured interviews. Regarding elicitation, firstly, a list of words by semantic domains was provided and then participants were asked what are the equivalent Western Yugur words to Mandarin Chinese and further words related to the semantic domain were elicited. Secondly, speakers were shown a picture of each word and asked to describe the picture in a short sentence. Thirdly, some group fieldwork meetings (3-4 people in each group) were organized for elicitation purposes. Fourthly, some words which are culturally significant have been elicited, by asking local people specific questions, such as does this word have a special meaning for Yugur people. Fifthly, Chinese and English dictionaries, especially some picture dictionaries, such as the Oxford Picture Dictionary (Parnwell, 1988), have also been used to elicit some of the words, such as daily life language and high-frequency words. Overall, the style of questions for the participants was open-ended and framed to gain a better understanding of Yugur language, terms, sounds and grammar.

Grouping words by semantic domains

Categorizing words into different ethnographic semantic domains can be useful for audiences who wish to extend their word knowledge in a given domain. Particular semantic domains such as kinship, food, animals, parts of plants, traditional clothing, color, numbers, religious words, *place and place names*, as well as ideophones, have cultural significance for the Yugur people, and seeing and hearing these words can help increase the motivation to revitalize the language. Some other examples include *directions, movement, describing words*, emotion words, *adverbs, question words, pronouns, suffixes, and so on*. There are many ways to elicit and group the words into different semantic domains or into an ethno-thesaurus. For example, the terminology of body parts was elicited by pointing to particular body parts, while anatomical diagrams were used to elicit the words for the internal organs (Bowern, 2015, p109). English or Chinese visual dictionaries also have been used to elicit the words according to different topics or thematic contexts, such as different colors.

Incorporating material from old word lists

Incorporating material from the existing wordlists is also vital but some of the words need to be examined and checked with the community speakers. The current new dictionary was partially drawn from a number of previous Western Yugur language research publications (Zongzhen Chen, 2004; Zongzhen Chen & Lei, 1985; Geng & Clark, 1992; Hermanns, 1951; Malov, 1957, 1967; Nugteren & Roos, 1996, 1998, 2003; Roos, 2000; Tenishev, 1976; J. Zhong, 2009), especially the existing Western Yugur-Chinese dictionary, which was compiled by Lei and Chen (1992). The future full-size dictionary will incorporate the previous works with due citation and acknowledgement.

Corpus

Dictionaries being produced today are more and more based on a large corpus of source materials. A corpus is a collection of written and spoken materials, which can be used for analysis of sounds, words, meanings, grammar and usage. Ideally, the dictionary entries should be selected from a corpus database gathered from fieldwork and existing documented publications. I have started a corpus for Western Yugur, and a future project is to build a much larger corpus of audio and video files as well as texts. It could be available to the public for online access and can be specifically used for teaching materials at schools or research. The existing body of texts, interviews and

stories will be incorporated into the corpus. They will be recorded in the original orthography and then have matching transcribed spelling in the practical orthography of Western Yugur. In this case, we can check to see that if all the words in the texts appeared in the dictionary, for example by doing a concordance.

Outcomes

This section discusses some experiences, methods and outcomes while compiling the actual trilingual dictionary. So far about 1300 dictionary entries have been created using Fieldworks Language Explorer (FLEx)¹, in a limited set of semantic domains, and some of the entries have richly detailed information, including pronunciation, examples of usage, grammatical information, and some aim to have associated cultural and ethnographic information drawn from my corpus and my native speaker knowledge. Some of the entries also contain sound files and pictures. This Western Yugur-Chinese-English trilingual online dictionary (trial version), has been presented through the Webonary website².

Development of a practical orthography

Dictionaries need orthographies. It is essential to design and develop a practical orthography with the speech community if there is not one already. Eastern and Western Yugur have been oral languages without written languages (although they used Old Uyghur script in the past). Since the 1950s, Chinese characters and Mandarin Chinese have become widely used amongst all Yugur people. Though IPA has been used for a few Yugur language publications, the community members, the primary targets of the dictionary, find it difficult to use.

Yugur community prefers to use the same script for Eastern and Western Yugur languages. After examining the sound system of the language, pinyin, the official phonetic system of Mandarin Chinese, combined with other Latin characters, was used to design a practical trial orthography. Mandarin is usually the second language for Yugur speakers, while people under 45 years old also have learnt pinyin at school. However, some Western and Eastern Yugur sounds are difficult to represent in pinyin. So several symbols from a few other Latin alphabets, including the Turkish practical orthography, have also been considered in this situation because Turkish is the most widely spoken Turkic language, and its spelling system is well established, and represents Turkish well. Nevertheless, people learn to read Mandarin characters and the pinyin romanization throughout China. This renders the modern Turkish alphabet less variable as orthographies for Western Yugur. Consequently, the proposed orthography uses a Latin alphabet, mainly based on the pinyin system, which accommodates not only regional variants, but can be extended to Eastern Yugur with some additional symbols. A practical orthography, which was designed and developed in consultation with the community, is used in this dictionary intended for widespread use. Using the community-developed orthography in language documentation, including the dictionary, makes these materials usable by Yugur community members.

Defining and translating the entries

Each entry of the dictionary aims to present the range of meanings and uses that the source language may have. Therefore, it is essential to discuss how the meaning of Western Yugur words should be defined. It is difficult to find real translation equivalents between two languages because

¹ <https://software.sil.org/fieldworks/>

² <https://westernyugur.webonary.org>

the culture and conceptual world of an endangered speech community can be very different from the target language community (Cablitz, 2011). The translation equivalents would be harder for compiling a trilingual dictionary since the two target languages are entirely different.

Defining or explaining the meaning of a word in the community language itself could help keep the original meaning of the word, which reflects native speakers' understanding of the word itself. Pawley (2011, p266) believes that most bilingual dictionaries are used as aids for translation purposes. However, the best monolingual dictionaries are closer to being ethnographic dictionaries, which are dictionaries that reflect native speakers' understanding and define complex lexical concepts using simpler vernacular lexemes. Elsewhere, some endangered speech communities have expressed a desire to keep their language with monolingual dictionaries with sizeable definitions (Corris, Manning, Poetsch, & Simpson, 2004, p43). Therefore, one of the ultimate goals for this dictionary project would be to create a monolingual dictionary with the community in the future.

While Pawley (2011, p266) concedes that translation of the lexical unit is still important, he argues that rich descriptions which give linguistic and cultural information are valuable to the speech community, and he proposes using analytic definitions which give a more precise and informative description than approximate translations. For the current trilingual dictionary, the first solution would be providing detailed analytic definitions, example sentences, grammar notes (such as if a verb is transitive and intransitive) for each entry, especially for core verbs and adjectives. See Figure 1 for an entry with glosses and analytic definitions.

The screenshot shows a dictionary entry for the word 'səq'. At the top, there is a blue header with a dropdown arrow on the left, the word 'Entry' in white, and a checkbox labeled 'Show Hidden Fields' which is checked. Below the header, the entry for 'səq' is displayed. It starts with the phonetic transcription ['su^hq] followed by speaker icons and the note 'cf: bas, yim₂. verb 1)'. The first definition is 'to squeeze out; to wring' with Chinese characters '挤; 挤出; 拧' and a note '(湿衣服)'. It provides a detailed English definition: 'To get liquid from something ([N-nəŋ]) (e.g. wet clothing, fruit) by pressing it firmly with one's fingers or hand.' This is followed by a Chinese translation: '通过用手指或手紧紧地挤压某物 (例如湿衣服, 水果) ([名词-nəŋ]) 挤出液体。' An example sentence in the community language is 'Go pornda gezga-nəŋ su-sə səq-oh^t-də.' with its English translation: 'He/she squeezed out the water from the clothing just then. 他/她刚才拧干了衣服上的水。' Semantic domains are listed as '7.3.2.7 - Take something out of something, 7 - Physical actions.' The second definition is 'to squeeze' with Chinese characters '挤; 挤压' and an English definition: 'To press something ([N-nə]) (usually a soft object, such as a soft toy which can make sounds) firmly together with one's fingers or hand.' The Chinese translation is '用手指或手按压某物 ([名词-nə]) (通常是柔软的物体, 如可以发出声音的玩具等)'. An example sentence is 'Go ghəzdar uzə-sə-nəŋ doghər oynash-nə səq-gha danju-də.' with its English translation: 'That girl kept squeezing her round toy. 那个女孩在不断地挤压她的圆形玩具。' Semantic domains are '7.7.4 - Press, 7 - Physical actions.' The third definition is 'to squeeze (eyes) shut' with Chinese characters '挤眼' and an English definition: 'squeeze (eyes) shut'. The Chinese translation is '挤眼'. Semantic domains are '3.5.6.3 - Facial expression, 7 - Physical actions.' The fourth definition is 'to wink; tip the wink' with Chinese characters '眨眼; 使眼色' and an English definition: 'wink; tip the wink'. The Chinese translation is '眨眼; 使眼色'. Semantic domains are '3.5.6.3 - Facial expression, 3.5.6.1 - Gesture, 7 - Physical actions.'

Figure 1 Defining a word (example I)

Some definitions provide syntactic frames for verbs. One important aspect for verb definitions is to give the information about what arguments the verb take and how they are expressed, for example we know that transitive verbs take objects, however the dictionary readers may want to know if a verb can take Accusative objects or sometimes also take Dative or Ablative arguments. Vernacular definitions in the style of the Collins COBUILD dictionary (Cobuild, 2009) simultaneously illustrate a syntactic frame of a word and its meaning. Such definitions would

show whether the verb takes an Accusative object or other arguments. Here below is an example to show the argument in the definitions with the verb *bas*.

bas	[ˈpas]	cf:	səq	verb	动词
-----	--------	-----	-----	------	----

1) to press; to press (down); to push (down) 按 ; 压 To purposely press or push (down) something [N-nə] hard or firmly against it with a hand, or with the body, sometimes with the foot. 用手或者用身体或脚, 有目的地用力 (向下) 按压某物[名词-nə] ◆ *Sen go arghash-nə haldər-gha bər bas*. (You) push the bag down towards the ground. 你把那个包往下压一压。

2) to hold firmly 压 ; 压制 To hold someone or something [N-nə] firmly under control, or to control something by holding it [bas-gha]. 控制[bas-gha]某人或某物[名词-nə]。 ◆ *Sar bər taghati-nə bas-gha yi-wat-də*. The eagle held the chick firmly (with its feet) and then ate the chick. 老鹰把小鸡按在身下吃掉了。

Figure 2 Defining a word (example II)

FLEx online dictionary (Webonary) and future work

The primary objective is to develop a well-designed Western Yugur-Chinese-English multilingual online dictionary and input words collected from fieldwork. Using FLEx it is easy and straightforward to publish an online dictionary, which is accessible to those who have access to the internet, and in particular, accessible to the primary audience, namely Western Yugur people who wish to learn or relearn their language. FLEx links fairly directly to Webonary, which can be updated relatively quickly. However, the current implementation of the dictionary in Webonary does not allow for easy use of audio or video. This deficiency in Webonary loses the potential advantage of online dictionary platforms, namely that video and audio should be easily included. These not only enrich the documentation, but they make the dictionary more valuable to the intended audience. For example, one effective way of learning the sounds of Western Yugur is to use an online dictionary or smartphone software apps.

Considering the different targeted audiences of the dictionary, future publications could be in different versions to suit multiple audiences of school students, adult learners of the speech communities and academics. At the same time, the lexical databases should be able to be reorganized and expanded, and thus serve as a base and source for other dictionaries and school materials. Furthermore, usability and accessibility testing of the dictionary should be carried out to obtain practical suggestions from the audience in order to improve the dictionary (Corris et al., 2004).

Crowdsourcing online dictionary

Nowadays, technology plays an important role in both language learning and language teaching. Based on the Webonary online dictionary, another aim is to create and develop a well-designed Western Yugur-Mandarin Chinese-English crowdsourcing online dictionary website and an app version for mobile devices. It is clear that social media and smartphone technology are penetrating

both cities and most of the remote areas of the Yugur community, and are becoming an additional way of communicating and sharing information. A potential advantage of future online dictionaries is the ability to crowdsource improvements and additions to the dictionary database.

Crowdsourcing will allow to add entries and related information volunteered by a large number of people through the internet, and create a new way of gathering words and looking up information. Native speakers and members of the community can participate by using the app and contribute words and related multimedia. The crowdsourcing dictionary could make use of the interactive capabilities of mobile devices through crowdsourcing to collect and refine new information. It should have functions which support adding new words by pronunciation saved into high-quality audio and uploading associated photos (high resolution photographs can be taken with any smartphone), adding additional senses, usages and example sentences of the words, and uploading further videos and other information, as well as a fuzzy search function. This could lead to increased engagement of speakers with their language communities and help revitalization, where, given a high level of participation and interaction within the Yugur community, the app could be a useful and important tool for sharing and preserving their languages. The community members can make further comments, discuss and make contributions to modify or add further entries, sounds, example sentences, pictures and videos. The online dictionary would be available to the public and offer easy access to local people. Eventually, it could be a significant-sized dictionary.

Discussion and Conclusion

The dictionary is primarily designed for the community, to allow speakers to maintain and pass on their language, and to allow learners easier access to the language. Dictionary-making has been guided by the principle that dictionaries should not be regarded simply as academic catalogues of words, but should be actively used within the communities, since ‘the dictionary is a critical reference for language communities involved with language teaching and learning’ (Hinton & Weigel, 2002). For example, a comprehensive and well-research dictionary that is easily understood by the local communities could be republished in parts for a children’s dictionary in schools.

There are lots of difficulties and challenges in compiling a comprehensive trilingual dictionary for an endangered language, especially an online dictionary. In the following are the key challenges. Firstly, we examined the phonetic system of the source language, and then developed a practical orthography with the community so the community can write in their own language. Secondly, we needed to solve the problems on how to define Western Yugur words. Defining or explaining the meaning of a word in the community language itself can help keep closer to the original meaning of the word. Analytic definitions were used in the target languages, which give a more precise and informative description than approximate translations. Lastly, there needs to be lots of time and probably funding to develop a good online dictionary, especially a crowdsourcing online dictionary, although currently we used Webonary as a free basic draft online dictionary.

The dictionary aims to provide a comprehensive list of Western Yugur words, grammatical information and examples of usage, as well as associated cultural, historical and ethnographic information. It attempts to make a good reference and encyclopaedic dictionary with ethnographic information. A significant task of language revitalization is spreading the understanding of the worth of keeping the language alive amongst speakers and non-speakers, while also building

motivation to maintain language use amongst speakers, and imparting their language to their children, grandchildren and interested community members. Here the use of the dictionary can be extended beyond providing linguistic information to recording culturally identifiable items, including cultural knowledge and customs in the related entries. In this way, the revival of cultural consciousness can be deepened by providing a source of awareness from culturally identifiable items in the original language. These items are a part of the cultural character and carry cultural feeling and mindset, but provide a more complete vision for cultural identity if they are embodied in the medium of the original language. Thus, learning from the ethnographic knowledge in the dictionary entries in the original languages can in turn motivate people to learn and maintain their language.

References

Atkins, B. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*: Oxford University Press.

Bowern, C. (2015). *Linguistic Fieldwork: A Practical Guide*: Springer.

Cablitz, G. H. (2011). Documenting Cultural Knowledge in Dictionaries of Endangered Languages. *International Journal of Lexicography*, 24(4), 446-462.

Chen, Z. (2004). *Xibu Yuguyu Yanjiu (Research into the Western Yugur Language)*. Beijing: China Minzu Photographic Art Press.

Chen, Z., & Lei, X. (1985). *Xibu Yuguyu Jianzhi (A Concise Grammar of Western Yugur)*. Beijing: Ethnic Publishing House.

Cobuild, C. (Ed.) (2009) Heinle Cengage Learning.

Corris, M., Manning, C., Poetsch, S., & Simpson, J. (2004). How Useful and Usable are Dictionaries for Speakers of Australian Indigenous Languages? *International Journal of Lexicography*, 17(1), 33-68.

Geng, S., & Clark, L. (1992). Sarig Yugur Materials. *Acta Orientalia Academiae Scientiarum Hungaricae*, 46(2/3), pp.189-224.

Hermanns, M. (1951). The Uigur and Angar Language in Kan su, China. *Journal of the Bombay Branch of the Royal Asiatic Society*, 26, pp.192-213.

Hinton, L., & Weigel, W. F. (2002). A Dictionary for Whom? Tensions between Academic and Nonacademic Functions of Bilingual Dictionaries. In W. Frawley, K. C. Hill, & P. Munro (Eds.), *Making dictionaries: preserving indigenous languages of the Americas* (pp. 155). Berkeley, CA: University of California Press.

Lei, X., & Chen, Z. (Eds.). (1992). Chengdu: Sichuan Minzu Press.

Malov, S. E. (1957). *Jazyk Zheltykh Ujgurov. Slovar' i Grammatika*. Alma-Ata: Izd-vo Akademii nauk Kazakhsko** SSR.

Malov, S. E. (1967). *Jazyk Zheltykh Ujgurov. Teksty i Perevody*. Moscow.

National Bureau of Statistics of China. (2011). Zhongguo 2010 Nian Renkou Pucha Ziliao - Quanguo Ge Minzu Fen Nianling, Xingbie de Renkou (Tabulation on the 2010 Population Census of the People's Republic of China: All Nationalities by Age and Sex). In: NBS, viewed 20 January 2016, <<http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm>>.

- Nugteren, H., & Roos, M. (1996). Common Vocabulary of the Western and Eastern Yugur Languages: The Turkic and Mongolic Loanwords. *Acta Orientalia Academiae Scientiarum Hungaricae*, 49(1/2), pp.25-91.
- Nugteren, H., & Roos, M. (1998). Common Vocabulary of the Western and Eastern Yugur Languages: The Tibetan Loanwords. *Studia etymologica Cracoviensia*, 3, pp.45-92.
- Nugteren, H., & Roos, M. (2003). Common Vocabulary of the Western and Eastern Yugur Languages: The Ethnonyms. *Rocznik Orientalistyczny*, 56, pp.133-143.
- Ogilvie, S. (2011). Linguistics, Lexicography, and the Revitalization of Endangered Languages. *International Journal of Lexicography*, 24(4), 389-404.
- Parnwell, E. C. (1988). *The New Oxford Picture Dictionary (Monolingual English Edition)*: Oxford University Press.
- Pawley, A. (2011). What Does It Take to Make an Ethnographic Dictionary? On the Treatment of Fish and Tree Names in Dictionaries of Oceanic Languages. *Documenting endangered languages: Achievements and perspectives*, 263-287.
- Roos, M. E. (2000). *The Western Yugur (Yellow Uygur) Language: Grammar, Texts, Vocabulary*. Leiden: University of Leiden.
- Tenishev, È. R. (1976). *Stroj Saryg-jugurskogo Jazyka (The structure of Saryg Uigur)*. Moscow: Nauka.
- Zhong, J. (2009). *Xibu Yuguyu Miaoxie Yanjiu (Descriptive Study of the Western Yugur Language)*. Beijing: Ethnic Publishing House.
- Zhong, Y. X. (2019). *Rescuing a Language from Extinction: Practical Steps with the Community for the Revitalisation of (Western) Yugur*. (Doctoral Dissertation), the Australian National University, Canberra.

A CRITICAL REVIEW OF THE INCLUSION OF ACRONYMS AND INITIALISMS IN *THE ENGLISH-CHINESE DICTIONARY*

Yongwei Gao

Fudan University

Abstract

Initialisms, often regarded as a subgroup of abbreviations, has been an integral part of the English vocabulary for several hundred years although the word *initialism* itself was first seen in use in 1899. The number of initialisms has been on a steady rise due to factors such as the ever-growing tendency of abbreviating technical terms and people's preference for lexical conciseness. *The Oxford English Dictionary* (OED)'s inclusion of initialisms such as *DVD*, *GHB*, *GTI*, *MDF*, *TGIF*, *VFX*, and *XXX* during its recent few updates is evidence enough of the increasing proportion of initialisms among English new words. However, the fact that the OED has only recorded a few hundred initialisms is indicative of the major problem in including this type of lexical items in dictionaries, namely failure to include frequently used initialisms. Many English-Chinese dictionaries are not immune to this.

The English-Chinese Dictionary (ECD), long considered to be one of the most authoritative English-Chinese dictionaries now available, leaves room for improvement when it comes to the inclusion of this type of abbreviations. Besides this, the dictionary has been plagued by two other major problems, namely the loose standard in including the initial forms of proper nouns and the profusion of initialisms that have already gone out of date. This paper will not only identify the problems found in the ECD, but also put forward suggestions for a better coverage of initialisms.

Key Words: initialisms, acronyms, abbreviations, *The English-Chinese Dictionary*

1. Overview

Acronym and *initialism* are defined by *The Oxford English Dictionary* (OED) as "a group of initial letters used as an abbreviation for a name or expression, each letter or part being pronounced separately" and "A word formed from the initial letters of other words or (occasionally) from the initial parts of syllables taken from other words, the whole being pronounced as a single word" respectively. According to *Merriam-Webster Collegiate Dictionary*, *acronym* dates back to 1943 while *initialism* 1899¹. Acronyms and initialisms, as a subgroup of abbreviations, has been an integral part of the English vocabulary since they began to appear in the first half of the 20th century. Frederic Stanley Dunn (1911: 130) described the emergence of initialisms in his article in *The Classical Weekly* in which he exemplified the use of a dozen college initialisms such as *BA*, *MA*, *LLD*, and *MD*. With the influx of acronyms and initialisms to the English vocabulary, dictionaries that aim to collect them were compiled in the second half of the previous century. Gale's *Acronyms, Initialisms and Abbreviations Dictionary* (AIAD) is a case in point. Its compiler wrote in the preface that "In modern times, breakneck progress in electronics, space exploration,

¹ Both dictionaries furnish *acronym* with a second sense, namely the synonym of *initialism*.

and data processing brought new concepts, new projects, and new instruments. It also brought new acronymic forms to save precious inches of newsprint and precious seconds of broadcast time, to serve as cloaks of military secrecy and as spotlights on products, ideas, and programs that the public was expected to support, admire, or purchase” (1960 vii). With an initial coverage of 12,000 entries in 1960, AIAD underwent frequent revisions, and in 1980, its seventh edition recorded 211,323 entries, an eighteen-fold increase in twenty years. Its latest edition, namely the 50th edition published in 2016, comprised of four volumes, recording over 500,000 entries.

As the use of initialisms is an indispensable part of dictionary compilation (e.g. used to indicate different languages in etymological information), English general dictionaries were usually eager to welcome them into their ranks of entries. Let’s take the fifth edition of *the Concise Oxford Dictionary* (COD) for instance. Abbreviations were not listed as separate entries in the body of the dictionary, and they were collected in one of its appendices. The appendix ran at least a dozen pages, including initialisms (e.g. AA [antiaircraft; Automobile Association], AP [Associated Press], FBI [Federal Bureau of Investigation], and HWM [high-water mark]) and abbreviations (e.g. *dept* [department], *inc.* [incorporated], and *pt* [part]) as well. However, when the sixth edition of the COD came out in 1976, the abbreviations were incorporated into the dictionary proper and many new initialisms were added, such as AA [Alcoholics Anonymous], ABC [American Broadcasting Company], BIS [Bank for International Settlements], and DA [deposit account]. Bilingual dictionaries in China, however, were rather slow in including acronyms and initialisms. Juanyun Zhang (1981: 173-174), though focusing her discussion on the inclusion of initialisms in Russian-Chinese dictionaries, pointed out that “Up to now, the number of acronyms and initialisms in today’s bilingual dictionaries is rather small, with the exception of dictionaries such as *A New English-Chinese Dictionary* (NECD)”. As the compilation of *The English-Chinese Dictionary* (ECD) began shortly after NECD’s publication and some of its compilers previously worked on NECD, the ECD could, to some extent, be considered a large-sized version of NECD. As a result, these two dictionaries adopted similar criteria in their inclusion of headwords. Long viewed as one of the most authoritative English-Chinese dictionaries now available, the ECD has been widely commended for its 220,000 entries, among which acronyms and initialisms account for a considerable percentage. Several hundred of them were added when the dictionary underwent a major revision from 2001 to 2007, such as AFV (alternative fuel vehicle; armored fighting vehicle²), DRM (digital rights management), HBA (host bus adapter), MOP (most outstanding player), SARS (severe acute respiratory syndrome), and SKU (stock-keeping unit). Nevertheless, due to factors such as the emergence of several hundred or even thousand new acronyms and initialisms in the English language in the past decade and the fact that some initialisms have fallen into disuse, the ECD leaves room for improvement when it comes to the inclusion of these abbreviations. This paper will not only identify the problems found in their inclusion in the ECD, but also put forward suggestions for a better coverage of them.

2. The inclusion of acronyms and initialisms in the ECD

When the ECD was first published in 1991, it boasted a total of 200,000 entries which almost covered the whole spectrum of the English vocabulary. Words, new or old, popular or obsolete, were recorded in the two-volume dictionary. Acronyms and initialisms were of no exception as they were given equal importance as other categories of lexical items. Their number exceed several thousand, and some of the examples are AAR (against all risks), BADGE (Base Air

² This is not a new initialism as it made its way into the COD as early as 1964.

Defense Ground Environment), *CCTV* (closed-circuit television), *HCG* (human chorionic gonadotrophin), *PNG* (persona non grata), *TKO* (technical knockout), and so on. Like many common English words, initialisms can be polysemous because they can be short for several words or phrases. The ECD took a rather inclusive policy in recording the various full names of some initialisms. *AA*, for example, was provided with nine senses, each representing a different word, phrase or organization, namely “achievement age”, “ack-ack”, “air-to-air”, “Alcoholics Anonymous”, “antiaircraft”, “antiaircraft artillery”, “associate in arts”, “author's alterations”, and “Automobile Association”.

The supplement to the ECD, which was published in 1999, recorded several dozen new initialisms, such as *ABS* (anti-lock braking system), *CDM* (cold dark matter), *CVS* (chorionic villus sampling), *HOV* (high-occupancy vehicle), *IPO* (initial public offering), and *MBO* (management buyout). The second edition of the ECD was published in 2007, with an addition of 20,000 new entries, among which several hundred were acronyms and initialisms. Some of these most frequently used ones include *BGH* (bovine growth hormone), *COPD* (chronic obstructive pulmonary disease), *DSL* (digital subscriber line), *EPS* (earning per share), *FMV* (full-motion video), *SWM* (single white male), etc. Some of them were relatively new as they were included by English monolingual dictionaries in the ensuing years, such as *CGT* (capital gains tax; OED: June, 2016), *FOI* (freedom of information; OED: Mar., 2017), *EVOO* (extra virgin olive oil; OED: Jan., 2018), *EBD* (emotional and behavioral difficulties/disorders; OED: Jan., 2018), *EEPROM* (electrically erasable programmable ROM; OED: Jan., 2018), and *GHB* (gamma-hydroxybutyrate; OED: June, 2018). There are at least a dozen acronyms and initialisms that have been not recorded by major English dictionaries such as the OED, Merriam-Webster, and *Collins English Dictionary* (CED), e.g. *CAGR* (compound annual growth rate), *PRICE* (protect, rest, ice, compression, and elevation), etc.

Of the several thousand acronyms and initialisms in the ECD, we can find some easily identifiable patterns:

- A. Most of them are technical terms which are subsumed under disciplines such as computing, medicine, chemistry, biochemistry, telecommunications, business, social sciences, education, and the military. Among these new additions, computing and medicine initialisms top the list, and some of the them are exemplified in Table 1:

Table 1 Computing and medicine initialisms

computer	full name	medicine	full name
ASP	application service provider	DNR	do not resuscitate
ISV	independent software vendor	DOT	directly observed therapy
PGP	pretty good privacy	HPS	hantavirus pulmonary syndrome
SIP	session initiation protocol	PDD	pervasive developmental disorder
SLA	service level agreement	PGD	pre-implantation genetic diagnosis
SSL	Secure Sockets Layer	VF	ventricular fibrillation

B. The number of acronyms is on a gradual rise although they still pale by comparison with initialisms. Relatively new acronyms recorded by the ECD include *ASBO* (anti-social behavior order), *BOGOF* (buy one get one free), *DWEM* (dead white European male), *ELISA* (enzyme-linked immunosorbent assay), *IMAP* (Internet Mail Access Protocol), *SARS* (severe acute respiratory syndrome), and so on. The tendency of creating acronyms imitating the sound of existing words is also palpable among the ECD's acronyms, as is evidenced by examples such as *GIFT* (gamete intrafallopian transfer), *INSET* (in-service education training), *MUD* (multi-user dungeon), *PRICE*, and *SAD* (seasonal affective disorder).

C. There is a growing trend of using the number 2 (two) to replace "to" in the making of new initialisms, all of which are related to online activities. The earliest example is undoubtedly *F2F* (or *f2f*) which stands for "face to face", a term popularly used in online chat. The supplement to the ECD, as the earliest English-Chinese dictionary to record this initialism, provided three illustrative examples for it, the earliest of which was dated 1993. The emergence of *B2B* (business to business) one year later popularized this trend, as is attested by several copycat words such as *B2C* (business to consumer), *B2E* (business to employee), *B2G* (business to government), *C2C* (consumer to consumer), *G2B* (government to business), and *P2P* (person to person or peer to peer).

D. The first edition of the ECD included many acronyms and initialisms that represent the names of organizations, and in the second edition more such abbreviations were added, such as *DARPA* (Defense Advanced Research Projects Agency), *DEA* (Drug Enforcement Administration), *FEMA* (Federal Emergency Management Agency), *MORI* (Market and Opinion Research International), *SFO* (Serious Fraud Office), and *SADC* (Southern African Development Community).

3. Problems in the inclusion of acronyms and initialisms

Upon perusal of the acronyms and initialisms recorded by the ECD, we can identify four major types of problems, namely the failure to record relatively recent acronyms and initialisms, their imbalanced inclusion and asymmetrical treatment, the profusion of initialisms that have already gone out of date, and the failure to draw a distinction between acronyms and initialisms.

3.1 The absence of relatively new acronyms and initialisms

The English vocabulary has been growing at a faster pace than before. Annually, there are several hundred or even thousand lexical additions in the language, among which there is a big proportion of acronyms and initialisms. In the past few decades, the number of acronyms and initialisms has been on a steady rise due to factors such as the ever-growing tendency of abbreviating technical terms and people's preference for lexical conciseness popularized by the advent of the Internet. In consequence, more and more acronyms and initialisms have been recorded by dictionaries of all kinds. *The Oxford Dictionary of New Words*, for example, included several dozen acronyms and initialisms such as *ADD* (attention deficit disorder), *BBS* (bulletin board system), *FOB* (friend of Bill), *MACHO* (massive astrophysical compact halo object), *NAFTA* (North American Free Trade Agreement), and *PC* (political correctness). *The Oxford English Dictionary* (OED), though sometimes regarded as rather initialism-unfriendly (for its niching of many initialisms and its non-inclusive policy), manages to record some of the latest acronyms and initialisms through its quarterly update, and it has so far included initialisms such as *BEE* (black economic empowerment), *DVLA* (Driver and Vehicle Licensing Agency), *GAAP* (generally accepted accounting principles), *GWOT* (global war on terror), *MDF* (medium density fiberboard), and *TGIF* (Thank God it's Friday) during its recent few updates. English-Chinese dictionaries did a better job in this regard. *NECD*, for example, not only recorded many new initialisms (e.g. *AP* [advanced placement], *ARM* [adjustable-rate mortgage], *BRIC* [Brazil, Russia, India and China], *CGI* [common gateway interface], *DNP* [did not play], *DVR* [digital video recorder], etc.) in the dictionary proper, but also collected about three hundred online abbreviations (e.g. *AFAIK* [as far as I know], *BCNU* [be seeing you], *FOAF* [friend of a friend], *GTG* [got to go], *MYOB* [mind your own business], *WTF* [what the fuck], etc.) in one of its appendices. *A New-era Dictionary of English New Words* (NED), fresh from the oven, includes 400 abbreviations (mostly initialisms) which account for almost 10% of all the entries in the dictionary. Some of these initialisms are *BDNF* (brain derived neurotrophic factor), *BF* (best friend), *CDO* (collateralized debt obligation), *DLC* (downloadable content), *FBS* (fetal bovine serum), *GMO* (genetically modified organism), and so on.

If we compare the acronyms and initialisms in the ECD with those new ones in both *NECD* and *NED*, we can find that the ECD had failed to include some of the abbreviations that were popularly used when it was being revised in the first few years of this century. *CBD* is one of such examples. The ECD does include *CBD* as an initialism, but it stands for *cash before delivery*. The one for *central business district* is conspicuously absent. The OED dates the initialism back to 1951 although the dictionary did not include it until June, 2016. Considering the fact that the ongoing revision of the OED is a gradual process, the dictionary's inclusion of *CBD* sixty years after its first use is understandable. Merriam-Webster, on the other hand, was quick to record its use as the word first appeared in its 11th edition published in 2003. Other initialisms that the ECD failed to record are shown in Table 2:

Table 2 Initialisms to be recorded in the ECD

Initialism	Full name	Date of first use
GMO	genetically modified organism	1989
LGBT	lesbian, gay, bisexual, and transgendered	1992
LSB	lower sideband	1971
OMG	oh my God	1917
PB	personal best	1971
PSA	public service announcement	

3.2 Imbalanced inclusion of acronyms and initialisms

Like ordinary words, many acronyms and initialisms may, to some extent, be related to one another as they may denote similar or even opposite concepts or things. The absence of one of two or more related acronyms or initialisms has been a perennial problem in dictionaries. Merriam-Webster, for instance, included *POTUS* (president of the United States) in its 11th edition, but its related term *FLOTUS* (First Lady of the United States) is absent from the dictionary. The OED is not immune from this problem. In January 2018, it included *CEO* (chief executive officer) and *CFO* (chief financial officer), and in September it added *CDO* (chief data officer) in its update. But as *CEO* has so far spawned at least a dozen popularly used initialisms (e.g. *CIO*, *COO*, *CSO*, *CTO*, etc.), the OED needs to step up its effort in covering them. In the same vein, the ECD has been plagued by this problem, as is attested by Table 3:

Table 3 Imbalanced inclusion of related terms

Absent from the ECD	Related terms in the ECD
AR (augmented reality)	VR (virtual reality)
BOGO (buy one, get one)	BOGOF (buy one, get one free)
DWAI (driving while ability impaired)	DUI (driving while intoxicated)
EAP (English for academic purposes)	ESP (English for Specific Purposes)
HEO (high Earth orbit)	LEO (low Earth orbit) MEO (medium Earth orbit)
LGBT (lesbian, gay, bisexual, and transgender)	GLBT (gay, lesbian, bisexual, or transgendered)

POD (print-on-demand)	VOD (video-on-demand)
PPC (pay-per-click)	PPV (pay-per-view)

3.3 Asymmetrical treatment of acronyms and initialisms

Sometimes acronyms and initialisms of the same category are treated differently in the ECD. These kinds of differences are chiefly manifested in the following two ways.

Firstly, the forms of initialisms are not treated in an identical way. Let's take *VIP* (very important person) for example. In most dictionaries, the initialism is capitalized without any full stop. However, in the ECD, the word is shown in three different forms, namely *VIP*, *V.I.P.*, and *Vip*, while many similar initialisms are provided with only one form. This problem can be partially attributed to the fact that the ECD based many of its headwords upon American dictionaries in the 1980s that preferred to use periods in an initialism. An initialism with the first letter capitalized has always been a UK tradition. Although such initialisms are used so in reality (e.g. *Defra* [Department for Environment, Food, and Rural Affairs]), they are all capitalized in dictionaries. Such inconsistency in word form will surely mislead users.

Secondly, the indication of the POS label is sometimes problematic. Most acronyms and initialisms are nouns, and sometimes they can be used as verbs after they experience semantic shifting. Should they be labelled as they are actually used or indicated with the universal "abbr." (short for abbreviation)? In most dictionaries, monolingual or otherwise, there is no uniform method in indicating this type of information. *VAT* (value-added tax) is a case in point. *Oxford Dictionary of English* (ODE) regards it as an abbreviation while OALD treats it as an uncountable noun. However, there does exist a consensus among lexicographers to provide specific POS labels if the said acronym or initialism can be used in at least two word classes (e.g. *FTP* [file transfer protocol] as a noun and a verb, *GT* [*gran turismo*] as an adjective and a noun, *ID* [identification] as a noun and a verb, etc.). In the ECD, although many acronyms and initialisms are indicated with *abbr.* and only a few of them have been furnished with the noun label, namely countable nouns such as *IOU* (I owe you), and *UFO* (unidentified flying object). However, there are some ordinary initialisms that should not be granted noun status, as is shown in the following examples:

DA *n.* 鸭尾巴式发型 [*duck's ass*的首字母缩合]

DDT *n.* 【化】双对氯苯基三氯乙烷，滴滴涕（商品名，一种杀虫剂）
[<D(IMETHYL) + D(IPHENYL) + T(RICHOROETHANE)]

FET *n.* =field effect transistor

GABA *n.* 【化】γ-氨基丁酸 [*gamma-aminobutyric acid*的首字母缩合]

GTP *n.* 【生化】三磷酸鸟苷 [<GUANOSINE TRIPHOSPHATE]

MAT *n.* 抑制微生物停留技术 [<Microbial Anti-attachment Technology]

3.4 The profusion of old-fashioned acronyms and initialisms

Due to reasons such as the change of names and technical obsolescence, some acronyms and initialisms may fall into disuse after they are recorded by dictionaries. *DfES* stands for Department

for Education and Skills which was a UK government department between 2001 and 2007. According to the 12th edition of CED, *CDV* stands for CD-video or compact video disc, and it had its moment. But now as people shift to watch videos in DVD or through online streaming, the word is no longer used. At least one hundred of acronyms and initialisms in the ECD will face the lexicographical chop because of being out-of-date. Table 4 lists some of them:

Table 4 Initialisms to be deleted from the ECD

initialisms	Full name	Notes
ADC	advanced developing countries	used in 1980s & 1990s
ATV	Associated Television	available 1955 - 1968
COL	computer-oriented language	obsolete
GARP	Global Atmospheric Research Program	1967 - 1982
HCS	human chorionic somatomammotrophin	Given way to human placental lactogen (hPL)
MLF	multilateral nuclear force	1960s
PMS	post-moshing syndrome	flash in the pan
SAC	Strategic Air Command	1946-1992

3.5 The failure to distinguish between acronyms and initialisms

The widely adopted criterion in drawing distinctions between acronyms and initialisms is to see whether the abbreviation in question is pronounced as a word or just letter by letter. But sometimes the line between them seems to be blurred as some abbreviations may be regarded as either acronyms or initialisms. Let's take *DAT* (digital audio tape) and *FOB* (fresh off the boat) for example. Although the etymological information in the OED indicates that both of them are acronyms, they are provided with two ways of pronunciation. Learner's dictionaries such as LDOCE, COBUILD, and Macmillan regard *DAT* as an acronym while bigger dictionaries such as Merriam-Webster, CED, and American Heritage simply treat it as an initialism. Also enjoying different treatment is *FEMA* (Federal Emergency Management Agency), an agency of the US Department of Homeland Security. American Heritage treats it as an acronym, and the online *Wiktionary* believes it is the homophone of *femur* while Oxford Dictionaries regards it as an initialism. Against this backdrop, it seems to be understandable for the ECD's occasional failure in telling them apart. Table 5 lists some of the acronyms in the ECD that were misidentified.

Table 5 Mistaken-identity abbreviations in the ECD

Acronym	Full name	Pronunciation
BOGOF	buy one get one free	/'bɒgɒf/
CAD	computer-aided design	/kæd/
CMOS	complementary metal oxide silicon	/'si:mɒs/
DEFRA	Department for Environment, Food, and Rural Affairs	/'defrə/
FSBO	for sale by owner	/'fɪzbəʊ/
IMAP	Internet Mail Access Protocol	/'ɪmæp/
MOAB	massive ordnance air blast	/'məʊæb/
NEPAD	New Partnership for Africa's Development	/'ni:pæd/
SKU	stock-keeping unit	/skju:/
SOCO	scene of crime officer	/'səʊkəʊ/

4. Suggestions for better coverage

As acronyms and initialisms come and go, only a certain proportion of them become a permanent part of the vocabulary. With hundreds of new acronyms and initialisms on hand each year, how should lexicographers record them? For the future revision of the ECD, I believe the editors should adopt a two-pronged approach to their inclusion.

The first prong concerns itself with the inclusion of new acronyms and initialisms on the basis of their frequency (either in the World Wide Web or news archives) and their currency (i.e. through Google News). *IoT* (Internet of things), appeared 244,000,000 times (as of May 6, 2019) if we search it only through Google and 56,200,000 times if we combine it with its full name. Such extremely high frequency guarantees a place for the word. For many other new acronyms and initialisms, they may only need 50,000 - 100,000 occurrences to make the cut. Some of the new batch of popular initialisms that should be included in the ECD are exemplified in Table 6:

Table 6 Popular initialisms to be included

Initialism	Full name
BPD	borderline personality disorder
CLV	customer lifetime value
CPM	cost per mille
DAU	daily active users
HIIT	high-intensity interval training
ICO	initial public offering
NDA	non-disclosure agreement
PUA	pickup artist

The use of Google News or other news websites presents us with a better picture of how often the searched item has been used recently. This method may of especial use if we want to weed out technical terms few laymen use. *MSCTA* (multi-slice computed tomography angiography), for instance, appeared 31,400 times in the WWW, but was used only once through Google News search. As a result, this medical term should be excluded.

The second prong involves the use of the so-called non-inclusion policy which dictates the following four categories of acronyms and initialisms should be excluded.

A. Slang words. Among new acronyms and initialisms, slang words account for a large proportion. The Internet has been proved to be a fertile ground for slangy acronyms and initialisms. Due to their slangy nature, this type of abbreviations is mostly used by in-groups and thus should be avoided. What should be mostly eschewed are vulgar Internet slang terms. *LMAO* (laughing my arse / ass off) is one of the lucky terms that have made their way into dictionaries such as Oxford Dictionaries and Merriam-Webster. However, many of its variants, such as *LMHO* (laughing my hiney off), *LMDO* (laughing my dick off), *LMNO* (laughing my nuts off), *LMTO* (laughing my tits / testicles off), and *LMVO* (laughing my vagina off), should be kept at bay. Other types of slang words that should not be made candidates for dictionary entries include military slang (e.g. *BMO* [black moving object] and *MFWIC* [mother fucker who's in charge]), subcultural slang (e.g. *ASD* [anti-slut defense]), *STP* [stand-to-pee], and *WGWAG* [white girls with Asian guys]), slangy online abbreviations (e.g. *JFGI* [just fucking Google it] and *STFW* [search the fucking Web]), and so on.

B. Analogic abbreviations. Sometimes a well-established initialism can have some variants due to alliterative factors. *TBA*, for example, can be normally interpreted as “to be announced” or “to be arranged”, but according to *Wiktionary*, the letter *A* also stands for *added*, *advised*, *aired*, *affirmed*, *answered*, *assigned*, and *assessed*. As these analogic initialisms may not be used as often as their prototype, they should be excluded. The same can be said of the variants of *TBD* (to be determined), such as “to be dated”, “to be decided”, “to be declared”, “to be defined”, “to be

deducted”, “to be delivered”, “to be designed”, “to be disclosed”, and “to be discussed”.

C. Overly technical terms. Initialisms abound in fields such as computing and medicine. Cheng (1994: 683) pointed out that “Acronyms are useful because they simplify, facilitate, and accelerate communication. They have become the shorthand of medicine. Physicians, especially cardiologists, like to use or invent these medical timesavers and are good at the task.” As a result, there are numerous medical acronyms and initialisms. *Dorland's Dictionary of Medical Acronyms and Abbreviations*, for instance, has a total number of more than 90,000 entries and definitions in its seventh edition. Stanley Jablonski, the original compiler of the dictionary, wrote in the preface to the first edition that “In medicine, they are used as a convenient shorthand in writing medical records, instructions, and prescriptions, and as space-saving devices in printed literature” (1987 preface). *Syndrome*, a common medical term, appears over a thousand times in the dictionary, and many of its initialisms are only known to medical professionals, such as *ALDS* (albinism-deafness syndrome), *AAMS* (acute aseptic meningitis syndrome), *HACS* (hyperactive child syndrome), *IFDS* (isolated follicle-stimulating hormone deficiency syndrome), *MCAS* (middle cerebral artery syndrome), and *NMSIDS* (near-miss sudden infant death syndrome). As a result, such initialisms should be kept out of the ECD.

D. Humorous words. Sometimes initialisms are created purely for humorous effects. H. Rider Haggard, an English writer of adventure fiction, created the initialism *SWMBO* (she who must be obeyed, jocularly referring to one’s wife or female partner) in his novel *She: A History of Adventure* (1886-1887). Although the word has been seen in use now and then, the flippancy indicated in its use accounts for its failure to be included in major monolingual English dictionaries. Thus the ECD should follow suit. Similarly used initialisms that should be shunned as well include *FBI* (female body inspector), *PhD* (permanent head damage), *PLOKTA* (press lots of keys to abort), *RAS* (redundant acronym syndrome), *TINLC* (there is no lumber cartel), etc.

5. Conclusion

Algeo (1975: 231-32) believes that English initialisms are easier to make than a word of any other category, letting every person be creative, but also secretive and exclusive. Indeed, abbreviating the first letters of a bunch of words does not involve much effort on the part of the users, therefore acronyms and initialisms are and will be created at a faster rate than before. It is an undeniable fact that more such abbreviations will make their way into dictionaries in the same way as other types of neologisms. However, due to their nature of being semantically transparent, relatively stricter criteria should be adopted in the inclusion of new ones. Meanwhile, deletion of old or obsolete ones should also be made part of dictionary revision.

References

- Acronyms, Initialisms & Abbreviations Dictionary*. (1960) Detroit: Gale Research Co.
- Algeo, John. (1975). The Acronym and Its Congeners. *The First LACUS Forum*, 1974. Ed. Adam and Valerie Makkai. Columbia, SC: Hornbeam Press, 217-34.
- Bonk, Mary Rose. (1999). *Acronyms, Initialisms & Abbreviations Dictionary, 26th Edition*. Detroit: The Gale Group.
- Cannon, Garland. (1989). Abbreviations and Acronyms in English Word-formation. *American Speech* 64(2): 99-127.

- Cheng, Tsung O. (1994). Acronymophilia: The Exponential Growth of the Use of Acronyms Should Be Resisted. *British Medical Journal*, 309: 683-684.
- Dorland's Dictionary of Medical Acronyms & Abbreviations, 7th Edition.* (2016). Philadelphia: Elsevier.
- Dunn, Frederic Stanley. (1911). An Apocalypse in Abbreviations. *The Classical Weekly*, 4 (17): 130-132.
- Gao, Yongwei. (2009). *A New English-Chinese Dictionary, 4th Edition.* Shanghai Yiwen Publishing House.
- Gao, Yongwei. (2019). *A New-era Dictionary of English New Words.* Beijing: The Commercial Press.
- Goldstein, Milton. (1963). *Dictionary of Modern Acronyms & Abbreviations.* Indianapolis: Howard W. Sams & Co., Inc.
- Jablonski, Stanley. (2005). *Dictionary of Medical Acronyms & Abbreviations, 5th Edition.* Philadelphia: Elsevier.
- Lu, Gusun. (1993). *The English-Chinese Dictionary.* Shanghai Yiwen Publishing House.
- Lu, Gusun et al. (1999). *The Supplement to The English-Chinese Dictionary.* Shanghai Yiwen Publishing House.
- Lu, Gusun. (2007). *The English-Chinese Dictionary, 2nd Edition.* Shanghai Yiwen Publishing House.
- Mathangwane, Joyce T. (2015). Abbreviations and Acronyms: The Case of Thalosi ya Medi ya Setswana. *Lexikos*, 25, 233-245.
- Mattia, Fioretta Benedetto. (1997). *Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols.* Amsterdam: Elsevier.
- Mattia, Fioretta Benedetto. (2003). *Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols, Second, Revised and Enlarged Edition.* Amsterdam: Elsevier.
- Oxford Dictionary of Abbreviations, Second Edition.* (1998). Oxford University Press.
- Zhang, Juanyun. (1981). A Discussion of Abbreviations in Bilingual Dictionaries. *Lexicographical Studies*, 1: 173-179.

HOW DO PEOPLE UNDERSTAND THE ‘AUTHORITY’ OF AN ENGLISH DICTIONARY? ANALYSIS OF PEOPLE’S REACTIONS TO THE INCLUSION OF NEW WORDS

Yuri Komuro
Chuo University

Abstract

By their nature, dictionaries carry authority, regardless of the intentions of those who actually write dictionaries, and the kind of authority attributed to dictionaries is an old issue in lexicography. This paper examines how people sometimes ascribe an ‘authority’ to English dictionaries that differs from that which dictionaries actually deserve by examining how people react to the news of a new word or word sense being added to a prestigious dictionary, such as the *Oxford English Dictionary* or *Merriam-Webster’s Collegiate Dictionary*, with reference to two words: *marriage* and *cisgender*. Before examining specific cases, an overview of the inclusion guidelines for general dictionaries is given as the basis of discussion. Modern lexicography offers practical inclusion guidelines, such as evidence, longevity, and lexicographical significance, which are supported by corpus data and other sources. The reactions of dictionary users were analysed according to the guidelines, and their lexicographical validity was discussed. The case of *marriage* clearly showed that people ascribed an incorrect form of authority to the dictionary, and the case of *cisgender* demonstrated that the inclusion guidelines could give rise to both approval and criticism.

Key Words: authority, inclusion criteria, new words, new word senses

Introduction

Dictionaries, by nature, carry authority, regardless of the intentions of those who actually write dictionaries, and the authority ascribed to dictionaries is an old issue in lexicography. Newspaper articles cite dictionaries, people get excited about new words in dictionaries, and even Supreme Court justices refer to both general and legal dictionaries (Rubin 2010, Brudney & Baum 2013). *Merriam-Webster’s Learner’s Dictionary* (LearnersDictionary.com) defines *authority* as follows:

“1. [noncount] the power to give orders or make decisions: the power or right to direct or control someone or something

Only department managers have the *authority* [=right, power] to change the schedule...

2. [noncount] **a:** the confident quality of someone who knows a lot about something or who is respected or obeyed by other people

... She spoke **with authority** [=authoritatively] about the history of the building.

b: a quality that makes something seem true or real

His sincerity added much more *authority* [=credibility] to the story...”

The authority that dictionaries possess should be that defined by **2a** (although it may be difficult to distinguish clearly between the first and second senses) as dictionaries provide *linguistic* information collected and carefully processed by specifically trained language specialists. However, some people seem to ascribe the first meaning of authority to dictionaries, and they express some doubt about, or even hostile criticism over, the inclusion of new lexical items. This paper examines how people react to the news of a new word or word sense being added to a prestigious dictionary such as the *Oxford English Dictionary (OED)* or *Merriam-Webster's Collegiate Dictionary*, with reference to two words, *marriage* and *cisgender*, which generated some controversy when a new sense or entry was made in a dictionary.

Method

This study used a qualitative method. First, before examining specific cases, an overview of the inclusion guidelines for general dictionaries is given as the basis of discussion, based on studies by two experienced practical lexicographers: Diamond (2016) based on his experience at Oxford Dictionaries, and Stamper (2018) on her experience at Merriam-Webster's. Then, two specific cases of inclusion are considered as examples that met public reaction: a new word sense of *marriage* being added to the 11th edition of *Merriam-Webster's Collegiate Dictionary* in 2003 and *cisgender* entering the *OED* as a new word in 2015. The comments and criticism regarding inclusion of those words are analysed and their validity discussed.

Results

Generally, new words and word senses enter a dictionary when it is proven that they have been used over a certain period of time in various genres. Diamond (2016) summarises the *OED* inclusion criteria, which are generally applicable to modern English dictionaries. He lists four factors: evidence (written records of different sources), longevity (record of over 10 years), naturalisation (being understood as an independent word without any explanation), and lexicographical significance (contribution to a more systematic description of the lexicon), along with other favourable factors, explaining that they are flexible and work in combination. Similarly, Stamper (2018: 99–101) writes, “A word has to meet three criteria for entry into most general dictionaries,” and gives “widespread use in print”, “a long shelf life”, and “meaningful use”, which correspond to the above-stated evidence, longevity, and lexicographical significance, respectively. Also note that lexicographers rely on corpus data and other sources to judge whether a word merits entry, and that their decisions are supported by statistics and linguistic evidence. There is no mention of cultural norms or attitudes as considerations.

Despite meeting the above-mentioned criteria, the suitability of a new sense of *marriage* for inclusion was criticised. In 2003, *Merriam-Webster's Collegiate Dictionary*, the 11th edition, updated its definition of *marriage* by adding the subsense “the state of being united to a person of the same sex in a relationship like that of a traditional marriage” to cover the meaning of *marriage* in *same-sex* or *gay marriage*, which were the most frequent collocations of *marriage*. As same-sex marriage had been a hotly discussed social subject, *Collegiate* was not the first to make this update. Before 2000, the *OED* had already added a usage note saying, “The term is now sometimes used with reference to long-term relationships between partners of the same sex,” to the original entry. However, 6 years after the *Collegiate's* revision, the company was suddenly attacked by write-in campaigns claiming that the definition is a disgrace, as gay marriage is not

legally or morally acceptable. The movement was started by a *World Net Daily* article criticising Merriam-Webster (Unruh 2009).

“One of the nation’s most prominent dictionary companies has resolved the argument over whether the term “marriage” should apply to same-sex duos or be reserved for the institution that has held families together for millennia: by simply writing a new definition.”

Here, a dictionary is described as if it had acted as an opinion leader.

When the word *cisgender* entered the *OED Online* in June 2015, the news made headlines in *The Independent* and was also covered by *The Guardian* later in the same month. *The Independent* article had a simple, descriptive title, “‘Cisgender’ has been added to the Oxford English Dictionary”; it explained the etymology, history, and increased use of the word, and reports welcoming comments by some activists (Green 2015). By contrast, *The Guardian* published an Opinion column: “OK, it’s in the Oxford English Dictionary – but do you know what ‘cis’ means?”, questioning the recognition of the word by the general public (Lees 2015). The article had received 1,119 comments as of 27 June 2015.

Discussion

First, the reaction of people toward the inclusion of a new sense of *marriage* reported in Stamper (2018) will be examined according to the inclusion criteria. Those who joined the write-in campaign against the update of *marriage* were provoked by the message: “What do you do when the dictionary does not support your definition of a word?” and refused to accept the inclusion *because* they personally disagree with the concept, which is completely irrelevant to whether the meaning in question should be recorded. Their strong opposition tells us that they wrongly see a sense of authority in the dictionary that is based on “the power to give orders or make decisions: the power or right to direct or control someone or something”.

Unruh (2015) describes Webster’s *Collegiate* as, “One of the nation’s most prominent dictionary companies,” in his harsh criticism; Merriam-Webster *is* a respected authority and is surely deserving of such an evaluation in terms of its lexicographical quality, but it is not its mandate to deny anyone’s thoughts or ideas, let alone “resolve” any political or religious argument over the legality of same-sex marriage.

In fact, the *Collegiate* seems to have made a neutral decision, to reflect how the word *marriage* has come to be used *not* by revising the original or traditional definition, but by adding the subsense. Simpson (2016: 55–56) states:

“The issue of how to handle same-sex marriage introduced a new attribute, which gradually gained in social importance, and gradually demanded to be noted. But the means of dealing with this change is different for smaller, desk-sized dictionaries... If a dictionary alters the basic definition of marriage to accommodate this new feature, then it falsely creates the impression that this broader meaning has coexisted since the early days: this is the technique that smaller dictionaries are often forced to follow, through lack of space.”

The *Collegiate*’s approach keeps the original definition, “the state of being united to a person of the opposite sex as husband or wife in a consensual and contractual relationship recognised by law”, to avoid the possible misinterpretation stated above and shows a new attribute of the word *marriage*. Stamper (2018: 241) analyses the strong reaction from the public as follows:

“They believe that if we make a change to the dictionary, then we have made a change to the language, and if we make a change to the language, then we also make a change to the culture around that language.”

Next, I analyse two articles on the word *cisgender*. *The Independent* article is more straightforward and lexicographer-friendly. Green (2015) writes:

“The compilers of the Oxford English Dictionary might have viewed it as just another word, but for people who have spent their lives fighting for the equal treatment of transgender people, it represents much more than that.”

“Just another word” shows that the writer understands the lexicographical criteria, and that the word *cisgender* has just fulfilled them and made its way into the dictionary, nothing more and nothing less. The attitudes or interpretations made by trans groups reported also go along with the inclusion policy of the *OED*:

“Trans groups welcomed the word’s inclusion in the dictionary, arguing that it proved that issues of gender identity were now becoming part of the wider public consciousness.”

“Its addition to the Oxford English Dictionary reflects the significant increase in discussion of gender diversity over recent years. We welcome the increased awareness of its meaning which this inclusion will bring.”

Both “the wider public consciousness” and “the significant increase in discussion of gender diversity over recent years” are tested on corpus data, and their appreciation is rooted in the linguistically proven state of the word.

On the other hand, Lees (2015) presents interesting challenges to the *OED*’s inclusion criteria by stating:

“... if cis gains wider currency, I’d be happy to use it more. You could argue that its inclusion in the Oxford English Dictionary is evidence that the use of cis has become mainstream. Hardly. Put it in, by all means, but I’d rather use words that don’t need to be looked up in dictionaries...”

She uses a criterion that she calls “the hair salon test”. If her hairdresser does not understand a word, it is not “a good word”. Here, *good* probably means understandable or reader-friendly, as she explains that she prefers to use “non-trans” over *cisgender* “because it is clearer to your average reader”. Her criticism is made on the basis of “evidence” or “widespread use in print”, and she simultaneously welcomes the inclusion of the word, recognising its “lexicographical significance”. The word *cisgender* fills a blank in the vocabulary, as there was no word opposite to *transgender*. Dictionaries record words when they are linguistically or lexicographically proved to merit an entry. A word entering a dictionary, however, does not mean that people are required to use that word.

So far, I have discussed the *lexicographical* validity of dictionary users’ reactions to the inclusion of new words and word senses. People’s resistance to the inclusion of the new word sense of *marriage* may seem to be based on the idea of cultural prescriptivism, which dictionaries are often accused of, but this is incorrect. There is enough lexicographical proof of current use of *marriage* in the sense of “the state of being united to a person of the same sex in a relationship like that of a traditional marriage”; indeed, same-sex marriage exists in some jurisdictions. Mugglestone (2016: 553) states:

“Inclusion ... is often seen as a prescriptive act of legitimization, as proof that a form has ‘really’ entered the language... The dictionary-maker is constructed as gate-keeper, momentarily opening up the ‘bastion’ to new members, irrespective of the fact that, in modern evidence-based lexicography, the process runs in precisely the opposite trajectory. Usage—the democracy of words—governs the decision to include or exclude a given word or sense.”

It is interesting that in the case of *cisgender* both approval and criticism are made more constructively. That may be partly because it was a new word entering a dictionary, and did not *change* a word that was already in existence (= a traditional idea).

Conclusion

I have illustrated how people’s opposition to the inclusion of some words is sometimes completely irrelevant from a lexicographical perspective and sometimes follows the trajectory of lexicographers, although they are not fully equipped to make objective decisions. There is no good or bad lexicographical judgement about words or word senses to include. Dictionaries do not aim to be policy makers or lawgivers when deciding on the inclusion of certain words.

Understandably, dictionary buyers or users are not aware of how decisions on inclusion are made, but where does their mistaken perception of a dictionary’s authority come from? Stamper (2018: 248) notes that dictionary companies promote the dictionary’s authority in order to sell them. The old image of major lexicographical figures like Dr. Johnson and Noah Webster may still affect people’s perception. Further study is needed to determine how dictionaries are viewed in society, and this probably differs from country to country.

References

- Brudney, J. J. & Baum, L. (2013) Oasis or mirage: The Supreme Court’s thirst for dictionaries in the Rehnquist and Roberts eras, 55 Wm. & Mary L. Rev. 483 (2013), Retrieved from <http://scholarship.law.wm.edu/wmlr/vol55/iss2/4>
- Diamond, G. (2016) Chapter 33: Making decisions about inclusion and exclusion. In Durbin, P. ed. *Oxford Handbook of Lexicography*. Oxford: Oxford University Press, 532-545.
- Green, C. (2015) ‘Cisgender’ has been added to the Oxford English Dictionary. *The Independent*. 25 June 2015. Retrieved from <https://www.independent.co.uk/incoming/cisgender-has-been-added-to-the-oxford-english-dictionary-10343354.html> (Accessed 2 May 2019)
- Lees, P. (2015) OK, it’s in the Oxford English Dictionary – but do you know what ‘cis’ means? *The Guardian*. 25 June 2015. Retrieved from <https://www.theguardian.com/commentisfree/2015/jun/25/cis-oxford-english-dictionary-gender-transgender-hair-salon-test>
- Mugglestone, L. (2016) Chapter 34: Description and prescription in dictionaries. In Durbin, P. ed. *Oxford Handbook of Lexicography*. Oxford: Oxford University Press, 546-560.
- New words notes. June 2015. Oxford English Dictionary. Retrieved from <https://public.oed.com/blog/june-2015-update-new-words-notes/>
- Rubin, P. A. (2010) War of the words: How courts can use dictionaries in accordance with textual its principles. *Duke Law Journal*, 60, 167-206.

Simpson, J. (2016) *The Word Detective: Searching for the Meaning of It All at the Oxford English Dictionary*. New York: Basic Books.

Stamper, K. (2018) *Word by Word*. New York: Pantheon Books.

Unruh, B. (2009) Webster's Dictionary redefines 'marriage'. *World Net Daily*. 17 Mar. 2009. Retrieved from <https://www.wnd.com/2009/03/91995/>

**A STUDY OF THE REPRESENTATION OF VERB-NOUN HETEROSEMY IN *THE*
*CONTEMPORARY CHINESE DICTIONARY (7TH ED.)***

Yushuang Dong

Renqiang Wang

Graduate School, Sichuan International Studies University, Chongqing China

Abstract

This study is concerned with the representation of verb-noun heterosemy based on *The Contemporary Chinese Dictionary* (7th ed.) and analyzes the relationship of word class labeling between Modern Chinese and Modern English. Word class categorization, especially the categorization of analytic languages, like Modern Chinese and Modern English, has always been a hot topic in linguistic research. The multifunctionality of lexemes mainly occurs among nouns, verbs and adjectives. However, to date there is a lack of empirical study of Chinese dictionaries from the level of (communal) language system. Therefore, from the perspective of the Two-level Word Class Categorization Theory, this study conducts an empirical study which aims to explore the overall situation of verb-noun heterosemy as well as its microstructure in *CCD7*. The result indicates that there are 1426 verb-noun lexemes in *CCD7*, accounting for 2.05% of the lexemes in the dictionary, which is approaching the 3% of the percentage of verb lexemes and noun lexemes in it, among the lexemes, 18082 lexemes belong to verb lexemes and 29567 lexemes belong to noun lexemes. In addition, the study finds that the number of the verb-noun lexemes of *CCD7* is slightly higher than that of the 5th and 6th editions, it, instead, is unexpectedly much lower than Modern English dictionary. There are inconsistencies among definitions, examples and word class labeling. This study shows empirically how the verb-noun lexemes of Modern Chinese relates to that of Modern English, also propounds certain proposals which can truly reflect the picture of verb-noun lexemes in Modern Chinese dictionary.

Key Words: *The Contemporary Chinese Dictionary* (7th ed.), verb-noun heterosemy, the Two-level Word Class Categorization Theory, word class labeling, principle of parsimony

1. Introduction

Word class categorization, especially in the analytic languages, such as Modern Chinese, Modern English and Modern Vietnamese, has always been a hot issue in linguistic research. Linguistic categorization such as word classes has been referred to as the study of “God particles” of language at the 36th annual meeting of the German Linguistics Congress, held in Marburg, Germany in 2014. Many international journals have also dedicated special issues to address word class categorization, such as *Studies in Language* (2008, 2017), *Theoretical Linguistics* (2012), *Linguistic Typology* (2016) and *Cognitive Linguistic Studies* (2018). Word class categorization in Modern Chinese has attracted even more attention (e.g. Guo, 2002; Hu, 1996; Lu, 2013; Lv, 1979; Ma, 1983; Shen, 2009, 2012; Wang, 2006).

However, multifunctionality of lexemes or heterosemy, as an ineluctable problem in the study of word class categorization, has been a thorny problem. Hu (1996, p. 215), for example, stated that the problem of multifunctionality of lexemes, especially the verb-noun heterosemous lexemes and adjective-noun heterosemous lexemes, has long perplexed the grammatical field. But to date there is a lack of empirical investigation of Chinese dictionaries at the level of the (communal) language system. Thus, this paper is intended to study the representation of verb-noun heterosemy in *The Contemporary Chinese Dictionary* (7th ed.) (hereafter *CCD7*) and to further explore the discrepancies between Modern Chinese and Modern English.

2. Background

Word classes are the cornerstone of linguistic models at a variety of levels of investigation (Beck, 2002, p. 11). The study of word classes, especially the study of multifunctionality of lexemes (or heterosemy), has been a long-stand theme over the past two millennia, which has never been interrupted. In below we present an overview of previous studies on multifunctionality of lexemes, verb-noun heterosemy and word class labeling in *The Contemporary Chinese Dictionary*.

2.1 Multifunctionality of Lexemes

Multifunctionality of lexemes (or heterosemy) is a widely discussed topic especially in analytic languages. If someone wants to judge whether a lexeme is a multicategory word, it means that “multifunctional lexical item can be shown to have the properties of more than one syntactic category; that is, it sometimes manifests the properties of one category and sometimes those of another one” (Lefebvre, 2001, p. 109), which is a general property of linguistic systems in general. The multifunctionality of lexemes enjoys a fair amount of attention in the literature (e.g. Bisang, 2011; Bloomfield, 1933; Enfield, 2006, 2015; Ježek & Ramat, 2009; Robert, 2004; Sasse, 1993; van Lier, 2017). Bloomfield (1933, p. 204), for example, regarded the phenomenon as “class cleavage” or “multiple class membership”, and the latter has been accepted by some scholars (e.g. Hudson, 2007; Robins, 1989).

In addition, the study of multifunctionality of lexemes has become shrouded in related terminologies. According to Ježek and Ramat (2009, p. 395), “transcategorization is a diachronic process consisting in a categorial shift of a lexical item without any superficial marking”. Robert

(2004, pp. 120-138) has spoken of “categorical flexibility” and “transcategorical morphemes”. He stated that syntactic and semantic flexibility is shown in synchrony by these transcategorical morphemes. On the other hand, “precategorization” (Sasse, 1993) and “precategoriality” (Bisang, 2011) have been discussed in relation to the above terms. Likewise, Beck (2002, p. 5) also indicated that “Recategorization is the acquisition by a word of the typical properties of another part of speech and decategorization refers to the loss of properties typical of a word’s own lexical class”. As for the term “underspecification”, Farrell (2001, pp. 112-115) presented a detailed semantic underspecification analysis in the framework of cognitive grammar, which indicated that semantics is the effective way to handle some potential difficult matters as polysemy and idiomaticity in syntactic level.

If a polysemous lexeme has more than one meaning, there is some significant overlaps in semantic content. “Multifunctionality” (Harris, 1946), “intercategorical polysemy” (Zawada, 2005), “heterosemy”, “polysemy” and “homonymy” (e.g. Enfield, 2006, 2015; Lichtenberk, 1991; Persson, 1986) emerged at a historic moment. The main differences among them are listed as follows:

Homonymy arises when two completely unrelated senses happen to be expressed by two different lexical items with nothing in common except their coincidental phonological form” (Persson, 1986, p. 469).

Polysemy normally means the association of distinct (but related) meanings with one and the same lexeme; Heterosemy refers to cases (within a single language) where two or more meanings or functions that are historically related, in the sense of deriving from the same ultimate source, are borne by reflexes of the common source element that belong in different morphosyntactic categories (Lichtenberk, 1991, p. 476).

Thus, heterosemy is a special case of polysemy, where the different but related meanings of a given morpheme are associated with distinct grammatical contexts (Enfield, 2006, p. 297). In other words, apart from the semantic relatedness, the historical basis is another common source to connect categorical members between polysemy and heterosemy. Instead, the distinction between polysemy and homonymy is closely associated with the concept of the lexemes.

Multifunctionality of lexemes often appears in isolating languages with little or no morphology, such as Chinese, or languages with rich morphology (Ježek, 2016, p. 47; Wang, 2014b, p. 54). In dealing with the multifunctionality of lexemes, Harris (1946, p. 165) proposed three kinds of class strategies: homonymy, heterosemy and adding a new word class. Aarts (2007, p. 225) denied the existence of multiple class membership yet conceded that gradience exists in syntax and supported the view of homonym according to semantic relation. However, the multiple class membership (heterosemy), as a generally accepted view, has always been studied by many scholars (e.g. Bloomfield, 1933, pp. 206-208; Herbst, 2010, p. 164; Quirk et al., 1985, p. 720).

As for Chinese, some scholars denied the existence of word classes, Ma (1898/1983), for example, proposed that “when a word has a specific functional usage in sentence, it will fail on word class in lexicon”. Li (1924, p. 24) said, “In general, the class of a word depends upon the sentence in which it occurs; separated from its sentence, it has no class”. By contrast, the mainstream view held by Chinese academics is that the multifunctionality of lexemes is an objective existence, many scholars adhere to the “principle of parsimony” which should be as few as possible. (e.g. Guo, 2002, p. 101; Lu, 2013, p. 57; Shen, 2009, p. 4; Zhu, 1982, p. 39). However, this brings a

series of problems that the given word class may fail on its syntactic function, which also leads to the dilemma for learners of Chinese as foreign language. Inevitably, heterosemy, as a phenomenon of transcategorial multifunctionality, often occurs in analytic languages (like Modern Chinese and Modern English) (Robins, 1989, p. 214). In terms of its complexity, the multifunctionality of lexemes has been one of the mainstream directions of linguistic research.

2.2 Studies on Verb-Noun Heterosemy

Following the definition of heterosemy, verb-noun heterosemy can be regarded as lexemes that have two word-classes categories of verb and noun in lexicon. It is worth noting that meaning correlation is a necessary condition for verb-noun heterosemy. In this way, a lexeme may have verbs and nouns yet each of them can be used in specific sentence. As English, the lexeme “swim” has two-word classes in *Oxford Advanced Learner’s Dictionary* (9th edition, hereafter *OALD9*). The labeling is shown in Figure 1.

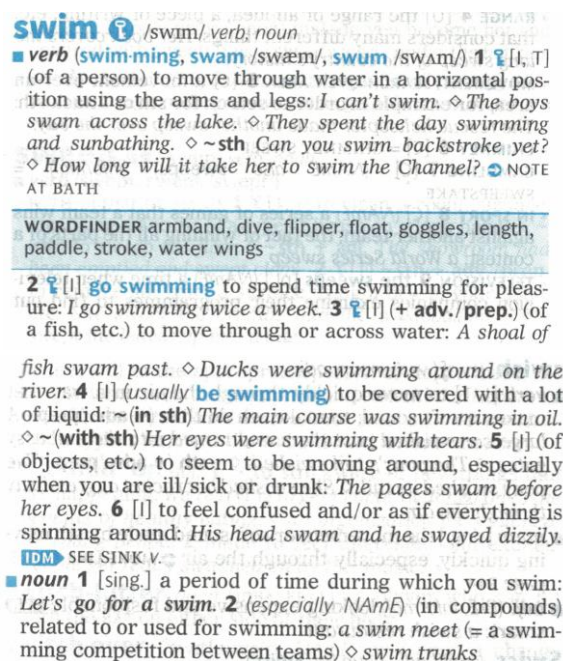


Figure 1. Screenshot of the entry for *swim* in *OALD9* (v. n.)

Relevant studies on verb-noun heterosemy might be categorized as follows. Firstly, the distinction between verbs and nouns is a controversial issue in the study of heterosemy, which has not been adequately addressed (cf. Baker, 2004; Jiang et al., 2011; Luuk, 2010; Rijkhoff, 2002; Shi, 2007). And the discussions fall on the question of nouny verb or nominalization (e.g. Dixon, 2005; Heyvaert, 2003; Hu, 1996; Lieber, 2016; Malchukov, 2006; Newmeyer, 1971; Zhu, 1982). Shen (2009) proposed that the Chinese content word class system is the “Inclusion Model”, which means nouns include verbs and verbs include adjectives. While Indo-European language is a “Discrete Model”, nouns and verbs are two separate categories. To be brief, a point which has been overlooked by Shen (2009) is that there is no clear-cut boundary between word token and word type. Ježek and Ramat (2009, p. 403) pointed out that “[i]f we start from the assumption that changes first happen in discourse and then enter the system through a diachronic process, we should distinguish between *faits de discours* (word token) and *faits de système* (word type)”.

What is clear is that the class membership of word type in the lexicon of communal language is its meaning potentials of word token in syntax at *langue* (Wang, 2014a, p. 346). We can see that the classification of verb-noun lexemes is the focus and difficulty in the study of heterosemy, more scholars study this issue from different perspectives (e.g. Conwell & Morgan, 2012; Hu, 2006; Luuk, 2010; Rijkhoff, 2002; Shi, 2007; Yao, 1980). Other scholars also discuss the relationship between verbs and nouns from the perspective of psychology and psycholinguistics (e.g. Caramazza & Hillis, 1991; Liu et al., 2011; Yang et al., 2002; Yi & Ni, 2018) and comparative linguistics (Jiang, 2012; Lu, 2012). In addition, scholars (Guo, 2002; Hu, 1996; Wang, 2016) have also investigated quantitatively Chinese word class. The focus, however, has been put on individual word (word token) categories, which does not explain the overall characteristics of the phenomenon of modern Chinese.

2.3 Word Class Labeling in The Contemporary Chinese Dictionary

The word class labeling of verb-noun lexemes in Modern Chinese has been a challenge for lexicographers (Lu, 1981, p. 151). *The Contemporary Chinese Dictionary* (hereafter *CCD*), since the 5th edition adds the labeling of word class, has inspired a new wave of study. More Chinese scholars study the problem of word class in the dictionary and make comparison with other dictionaries that label lexemes with word classes (e.g. Fang, 2010; Jiang et al., 2011; Wang, 2010, 2011, 2013; R. Q. Wang & Wang, 2016; Xu & Tan, 2006; Yang, 2016; Yang & Wang, 2018; Zhao, 2015).

The Contemporary Chinese Dictionary (5th ed.) (hereafter *CCD5*) is a milestone in the history of Chinese lexicography, it is the first time to complete the full labeling of word class in 2005. In order to emphasize the importance of this work, Xu and Tan (2006) elaborated part of speech tagging in *CCD5* through three sections. The first is an introduction about the system of part of speech adopted in *CCD* and the grammatical features of each part of speech; The second one is to demonstrate that *CCD* labels word class for monosyllables and multi-syllables on the basis of distinguishing words from non-words. The third part offers solutions to some divergence and provides reasons for doing so. In addition, Wang (2010, 2011, 2013) explored Modern Chinese word class systematically and comprehensively, not only for word class system validity, wordhood in Modern Chinese, but the whole heterosemous phenomenon. Wang and Zhou (2015) also proved that there is the relationship between heterosemy and frequency, and *CCD5* minimizes the number of multicategory lexemes artificially by following the principle of parsimony. From the perspective of the quantitative approach, Wang (2016) reported on an investigation of parts of speech based on *CCD5*, the study made a first attempt to plot the polyfunctionality distributions of individual parts of speech in Ord's system, which surprisingly shows approximately a hyperbola. As an updated version, *CCD6* has made a number of improvements compared to its predecessors. Zhao (2015) discussed five major improvements the *CCD6* has made as against the *CCD5* and four weak points in detail. Yang (2016) also compared the word class labeling of heterosemy between *CCD6* and *CCD5*. Yang and Wang (2018) made a systematic study of the representation strategies of multi-category lexemes from the perspective of the Two-Level Word Class Categorization Theory.

CCD7 is published in 2006, up to now, the relevant studies on *CCD7* is mainly the study of newly-increased words and the study of strategies for the treatment of self-reference and partial self-reference (Yang, 2019; Tan, 2018). These studies highlight the need for an in-depth analysis of

word class labeling of dictionary from the perspective of the synthesis of ontological and methodological study. To be specific, there is a lack of systematic and in-depth study on the word class labeling of verb-noun heterosemy within *CCD7*. In view of this, the study is intended to scrutinize the verb-noun heterosemy of *CCD7* from the perspective of the Two-level Word Class Categorization Theory in order to provide detailed account in three aspects: what (the status quo of word class labeling of verb-noun heterosemy in *CCD7*), why (the potential reasons behind this problem) and how (the importance of theoretical basis).

3. Methodology

3.1 Research Questions

Taking these studies as the point of departure, the overall aim of this study is to investigate the representation of verb-noun heterosemy using examples taken from *CCD7* and explore the diachronic comparison with other editions. In particular, the study is interested in contributing to the identification of microstructure, to help characterize the relation among word class labeling, definitions and examples, to explore the question of whether there exist differences between Modern Chinese and Modern English, to help identifying the proper theory for word class labeling. To be more specific, the following research questions are addressed by the study.

- (1) How does verb-noun heterosemy is represented in *CCD5*, *CCD6* and *CCD7*?
- (2) Are there any differences in word class labeling of verb-noun heterosemy between Modern Chinese and Modern English?
- (3) What are the possible reasons behind different types of word class labeling, and whether the Two-level Word Class Categorization Theory can provide a more reasonable labeling of verb-noun heterosemy?

3.2 Theoretical Basis

The Two-level Word Class Categorization Theory (TLWCCT), based on a complex adaptive system (CAS) (Beckner et al., 2009; Bybee, 2010) and cross-language nature of word classes (Croft, 1991; Croft & van Lier, 2012), is the new word class categorization model which aims to deal with the problems of analytic languages from the level of the communal language (Wang, 2014a). There are five pair of contents involved in this theory:

- (1) Two levels: The class membership of a word token in syntax at *parole* and the class membership of a word type or lexemes in lexicon at *langue*.
- (2) The differences between the two levels: The word class categorization of a word token in syntax is the expressional process of speaker's propositional act function (such as reference, statement and modification), it also can be judged by the syntactic function of word token and the relevance-markedness patterns of their meaning events or contexts; The word class categorization of a word type in lexicon is a self-organizing process of communal language, the core of which is the conventionalization.
- (3) The relation between the two levels: The class membership of a word type neither exists *a priori* nor is precatagorical, but evolves with the repeated use of word token in syntax at *parole*

(i.e. in various propositional speech act constructions), and the multifunctionally or multiple class membership of a word type is closely associated with the frequency of use (the type frequency and token frequency).

(4) The judgement of word classes: The class multiple (single-category or multi-category) of a word type is its meaning potential at *langue* and the conventionalized propositional speech act function embodied in the syntax, the conventionalized propositional speech act function needs descriptive linguists to investigate the propositional speech act function (the usage patterns) of a word token in syntax through the corpus-based sample analysis. It is also the categorization process in the degree of the propositional speech act function's conventionalization, in which the word class labeling of language dictionary is a typical representative.

(5) The criteria for conventionalization: Based on the corpus-based usage pattern, there are four criteria to judge conventionalization: token frequency, type frequency, time span and register variation (Wang & Chen, 2014), among which token frequency and type frequency are the main criteria to judge the lexemes.

3.3 The Dictionaries and Corpus Used in this Study

This study mainly discusses the verb-noun heterosemous lexemes on the basis of “the DIY Word Class Labeling Database of *CCD7*”. *CCD*, as a medium-sized dictionary, is the most authoritative Chinese dictionary in Modern China, whose status is similar to *Oxford English Dictionary* in the English World. The purpose is to standardize the words, popularize Beijing pronunciation and promote Chinese standardization. In 1978, the Chinese Social Science Academy published the official Chinese dictionary *Xiandai Hanyu Cidian (The Contemporary Chinese Dictionary)* by The Commercial Press. Since its inception, six editions were published respectively in 1983, 1996, 2002, 2005, 2012, 2016. A momentous revolution happens in 2005—the 5th edition comprehensively labels senses with word classes, which starts to distinguish the correlation between words and non-words, and the sixth and the seventh edition modifies it.

Crystal (2010, p. 303) pointed out that “from a typological viewpoint, English is in fact more similar to an isolating language like Chinese than Latin: there are few inflectional endings, and word-order changes are the basis of the grammar”. On the other hand, in light of the equivalent sum of word collection between *CCD7* and *Oxford Advanced Learner's Dictionary* (9th edition) (hereafter *OALD9*), this study resorts to the usage pattern of verb-noun heterosemy in *OALD9* and analyzes the differences of word class labeling on verb-noun heterosemous lexemes between Modern Chinese and Modern English.

This study selects “the Chinese National Corpus”^① (hereafter CN corpus), which is a large-scale balanced corpus with a wide range of material categories and a longtime span. Based on the advantages of word segmentation and part of speech tagging, the corpus can be retrieved by word and word class.

3.4 Procedures

Based on the approach of qualitative and quantitative, comparison and analysis, this study, generally, firstly inputs every entry into computer and builds the “DIY Word Class Labeling Database of *CCD7*”, then calculates the number of verb lexemes, noun lexemes and verb-noun

lexemes through the repeated siftings. In the last step, the study makes a diachronic comparison on the number and proportion of verb-noun heterosemy among *CCD5*, *CCD6* and *CCD7*. Synchronically, the study figures out the number of the whole dictionary entries and verb-noun lexemes in *OALD9*, and regards the lexemes as the comparative material, which aims at comparing the differences between Modern Chinese and Modern English. The study also investigates the relationships among definitions, examples and word classes labeling of verb-noun heterosemy.

After data collection, the data analysis is of more importance. A further step is carried by the study separates the verb-noun heterosemy from whole verb-noun lexemes in a narrow sense. In this study, only the verb-noun heterosemy which is labeled all senses with word classes can be chose. For instance, the entry 音 ('sound') is a verb-noun lexemes with partially word class labeling, the second and third sense are morphemes. Thus, it is not the object of the study. However, the entry 将军 (lit, 'general') is a verb-noun lexeme with fully word class labeling, it is included in the study. The detailed word class labeling is represented as follows:

音 yīn ① 图 声音;读音: ~律|~乐|口~|乐~|杂~|把这个字的~读准。② 消息: 佳~|~信。③ 指音节: 单~词|复~词。④ 图 读(某音): “区”字做姓时~欧。⑤ (Yīn) 图 姓。

Figure 2. Screenshots of the entry for 音 in *CCD7* (v. n.)

【将军】 jiāngjūn ① 图 将(jiàng)级军官。② 图 泛指高级将领。③ 图 大黄的旧称。④ (-//-) 图 将⑤。⑤ (-//-) 图 比喻给人出难题,使人为难: 他当众将了我一军,要我表演舞蹈。

Figure 3. Screenshots of the entry for 将军 in *CCD7* (v. n.)

After calculating the number and percentage of verb-noun lexemes, the conclusion will explain the representation of verb-noun heterosemy in *CCD7* and try to comb the problem of word class. In addition, it will conduct a case study on disyllable lexeme 斗争 ('struggle') on the basis of the CN corpus, aiming to find out whether there is a difference in the word class labeling of verb-noun heterosemy between Modern Chinese dictionary and Modern English dictionary.

4. Results

4.1 The General Description of Verb Lexemes, Noun Lexemes and Verb-Noun Lexemes in *CCD7*

There are total 69674 entries in *CCD7*, with 3769 multicategory entries and 3396 two-categories entries as well. Since *CCD7* fully labels senses with word class based on the distinction of words and non-word. We further classify the entries in terms of those fully labeled lexemes and partially

labeled lexemes. The results show that there are 1426 verb-noun lexemes, 18082 verb lexemes^② and 29567 noun lexemes with fully word class labeling.

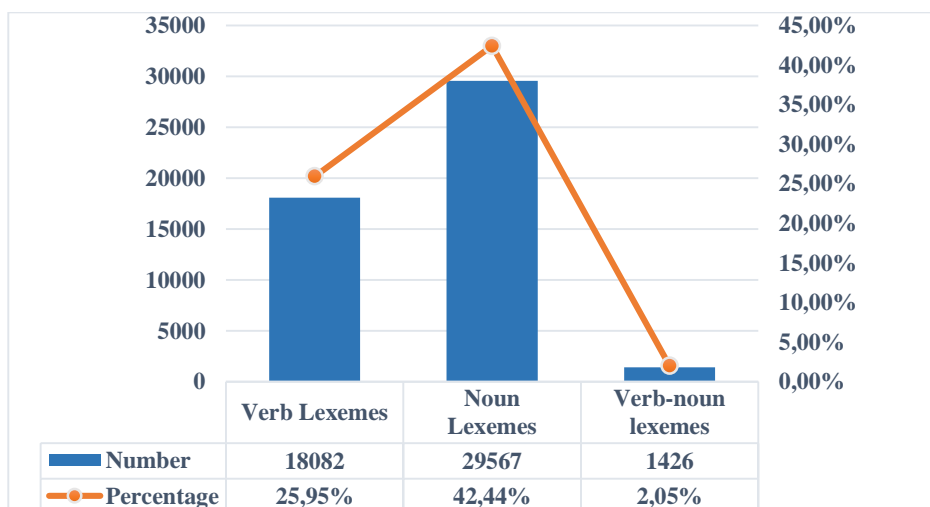


Figure 4. The distribution of verb lexemes, noun lexemes and verb-noun lexemes in *CCD7*

4.2 Comparative Description of Verb-Noun Lexemes between Modern Chinese and Modern English

4.2.1 Diachronic Description of Verb-Noun Lexemes among *CCD5*, *CCD6* and *CCD7*

Based on the self-built word class labeling databases (Wang, 2013; Yang & Wang, 2018), this study analyzes verb-noun heterosemy in three dictionaries respectively. The result shows that the number of *CCD7* has slightly increased in overall situation, of which 3015 lexemes are the multicategory lexemes, 2866 of them as two-category lexemes with fully word class labeling. Compared with *CCD5*, *CCD6* has distinct variations since the 6th edition revises many lexemes and adds nearly 3000 neologisms. The data in detail is presented in Table 1:

Table 1

Comparative Data of Verb-noun Heterosemy among CCD5, CCD6 and CCD7

	<i>CCD5</i>	<i>CCD6</i>	<i>CCD7</i>
Total Entries	65815	69232	69674
Multicategory Lexemes	2778	2997	3015
Two-category Lexemes	2649	2848	2866
V-N Heterosemous Lexemes	1311	1420	1426

As shown in Table 2, verb-noun heterosemous lexemes in the two-category lexemes account for 49.76% in *CCD7*, there is about a 0.1% decrease in that of *CCD6* (accounting for 49.86%). The verb-noun heterosemous lexemes of *CCD5* occupy 49.49% in the two-category lexemes. Thus, it can be seen that the small reduction in the percentages of verb-noun lexemes is not directly proportional to the sum increase in the whole entries.

Table 2

The Percentage of Verb-noun Heterosemy among CCD5, CCD6 and CCD7

Type	CCD5		CCD6		CCD7	
	Total Entries	Two-Category Lexemes	Total Entries	Two-Category Lexemes	Total Entries	Two-Category Lexemes
Number	65815	2649	69232	2848	69674	2866
Percentage	1.99%	49.49%	2.05%	49.86%	2.05%	49.76%

4.2.2 Synchronic Description of Verb-Noun Lexemes between CCD7 and OALD9

Here are total 47679 entries in *OALD9*, among which 4855 of them are handled as multicategories lexemes, 4402 of them as two-categories lexemes. In this dictionary, there are 2662 verb-noun lexemes in fully labeling, accounting for 60.47% in two-category lexemes, it, besides, occupies half of multicategory lexemes. This shows that the heterosemy is the widespread existence in Modern English dictionary. The result is shown in Table 3.

Table 3

The Percentage of Verb-Noun Heterosemy in OALD9

Type	Total Entries	Multicategory Lexemes	Two-category Lexemes
Number	47679	4855	4402
Percentage	5.58%	54.83%	60.47%

4.3 The Results of Word Class Labeling of Verb-Noun Heterosemy in CCD7

Word class labeling not only influences the overall situation of the dictionaries, but also the sense arrangement, definition and example at the internal microlevel (Wang 2006: 59). In view of this, we analyze the distributions of definitions and examples of verb-noun heterosemous lexemes in *CCD7*.

4.3.1 The Inconsistency between Definitions and Word Classes

Firstly, after authenticating the word classes of all the verb-noun lexemes, we find that although some lexemes are labeled both in verbs and nouns, definitions of them are not consistent with the word classes. The result is shown in Table 4.

Table 4

The Inconsistency between Definitions and Word Classes in CCD7

Number	Lexemes	Word Classes and Definitions
1	编导	动 编导和导演.
2	雕塑	动 造型艺术的一种,用竹木、玉石、金属、石膏、泥土等材料雕刻或是塑造各种艺术形象.
3	罚金	动 审判机关强制被判刑人在一定期限内缴纳一定数额的钱的刑罚。是一种附加刑,也可以独立使用.
4	跳绳	动 一种体育活动或儿童游戏,把绳子挥舞成圆圈,人趁绳子近地时跳过去.

4.3.2 The Inconsistency between Examples and Word Classes

Similarly, Table 5 shows that there are some uncoordinated relationships between examples and word class labeling.

Table 5

The Inconsistency between Examples and Word Classes in CCD7

Number	Lexemes	Word Classes and Examples
1	编导	动 编导和导演: ~人员.
2	采编	动 采访并编辑: 电视台的~人员.
3	处方	动 医生给病人开处方: 不是医生,没有~权.
4	代表	名 受委托或指派代替个人,团队、政府办事或表达意见的人: 全权~.
5	点评	动 评点,评论: 最后专家进行了精彩的~.
6	忌讳	动 对某些可能产生不利后果的事力求避免: 在学习上,最~的是有始无终.

7	交易	动 买卖商品：进行公平~.
8	热望	动 热烈盼望：~的目光.
9	特护	动 (对重病人) 进行特殊护理：经过几天的~，他终于脱险了.

5. Discussion

5.1 The Representation of Verb-Noun Heterosemy in CCD7

In the study of Modern Chinese grammar and Chinese dictionary word class labeling, heterosemy is a long-term problem that has plagued Chinese grammatical scholars and lexicographers. The verb-noun heterosemy occupies the largest proportion in the multicategory lexemes, and the accuracy of verb-noun heterosemy has a direct impact on the processing of heterosemy in dictionary. From Table 1, it is clear that the sum of dictionary entries is increasing. *CCD7* has a total of 69674 entries, whereas *CCD5* and *CCD6* have 65815 and 69232 entries respectively. From Table 2, the proportion of verb-noun heterosemy in two-categories lexemes has reduced slightly, which is a 0.1% decrease in that of *CCD6* and 0.27% increase from that of *CCD5*. Zhang and Yong (2007, p. 353-366) pointed out that dictionary compilation in the macrostructure should follow the rules and norms, and as to microstructure, accurate definitions, reliable example and outstanding function all should be kept consistently. Therefore, the quality of lexicography lies in the joint effect of macrostructure and microstructure. While the obstacle factor is still the judgement of heterosemy. The sum of the whole dictionary entry is equivalent between *CCD7* and *OALD9*, but the number of verb-noun heterosemy in *OALD9* is 2662, accounting for 5.58% in the whole dictionary, 54.83% in multi-categories lexemes and 60.47% in two-category lexemes, the latter is the 10.71% increase than that of *CCD7*. Therefore, as a typical analytic language, the number of verb-noun heterosemy of Modern Chinese is smaller than that of Modern English. To be brief, although the number and proportion of verb-noun heterosemy in the same kind of dictionaries have increased slightly, the number of verb-noun heterosemy in *CCD7* is far from enough compared with Modern English.

Dictionary microstructure mainly refers to the information organization structure inside the entry, including all the information behind the word head in the content (Zhang & Yong, 2007, p. 59). Chen and Huang (1993, p. 29) also pointed out that compiling entries mainly includes reasonable words, accurate definitions, concise style, accurate annotation, typical examples and scientific arrangement, which hold true for any kinds of dictionaries. Instead, *CCD7*, as the second updated version, also has some weak points in places. As Table 4 and 5 demonstrate, there are something incongruous among the definitions, examples and word classes. The word class labeling of some entries is still controversial. For instance, 跳绳 (lit, 'skip rope') is labeled as a verb, but it defines as 一种体育活动或儿童游戏 ('a sports activity or children's game'), which is the typical definition of nominal usage. Therefore, it is better to label this entry as a noun. In addition, some examples do not match the entries in their word classes. Take the entry 点评 ('comment') as the

example, it is marked as a verb and defined as 评点/评论 ('comment on/discuss'). However, in its specific use, 最后专家进行了精彩的点评 (lit, 'Finally, the expert made a wonderful comment'), the word 点评 in here is a noun, rather than a verb. In fact, this example represents the nominal use of this word.

5.2 Analysis of Reasons behind Different Types of Word Class Labeling

The reason why the word class categorization of analytic languages has become the "Godbach conjecture" lies in the mismatch of the nature of language and research emphasis (Wang & Zhou, 2015, p. 67). Wang (2013, p. 14) found that *CCD5* has the 2778 multicategory lexemes of the 51469 lexemes which are fully labeled senses with word class in whole entries, accounting for 5.40%, the 1311 verb-noun heterosemous lexemes account for 49.49% of the 2649 two-category lexemes. However, *OALD7* has the 4861 multicategory lexemes of the 46380 lexemes which are fully labeled senses with word class in whole entries, accounting for 10.48%, there are 25 types of multicategory, of which the 2643 verb-noun heterosemous lexemes account for 59.65% of the 4431 two-category lexemes (Wang, 2014b). In this study, for *OALD9*, the number of verb-noun heterosemous lexemes is 2662, accounting for 5.58% of the whole entries and 60.47% of the 4402 two-category lexemes. *CCD7* has 1426 verb-noun heterosemous lexemes, which occupy 2.05% of the whole entries. It is clear that the amount of word-collection of Modern English dictionary is smaller than that of Modern Chinese dictionary. On the other hand, however, the number of verb-noun heterosemous lexemes of Modern English dictionary is more than 1236 for that of the Modern Chinese dictionary.

Since the multifunctionality of lexemes in the communal language system is universal in the analytic languages, Modern Chinese and Modern English are typical analytic languages, the analytic feature of Modern Chinese should be higher than that of Modern English (Crystal, 2010, p. 303). Consequently, a reasonable result is that the proportion of Modern Chinese should also be higher than that of Modern English, only in this way can they accord with the accurate situation in Modern Chinese. Pitifully, that is so far from the truth in terms of comparison between *CCD7* and *OALD9*. Why does the Modern Chinese dictionary have such a lower number of verb-noun heterosemy? The main reason is that scholars abide by the principle of parsimony (the principle of fewest possible the multicategory words), which has been regarded as the golden rule in Modern Chinese grammar research. Modern Chinese concerns that the establishment of heterosemous usage will occupy the length of the dictionary and increase the cost of dictionaries compilation. Nevertheless, blindly following this principle may cause some problems in word class labeling of dictionary. The number of heterosemous lexemes is reduced artificially, which might be difficult to justify theory and might cover up the truth of language use.

As the main supporter in the principle of parsimony, Shen (2009) proposed the "Inclusive Model" of Chinese content words, which hold that verbs are included in nouns. In his opinion, two levels between the syntax and the lexicon do not exist. Based on this, we further calculate the percentage of verb lexemes, noun lexemes and verb-noun lexemes in *CCD7* built on the result of Figure 1. The data is shown in Figure 5. It is found that there are total 49075 lexemes^③, of which the 29567 noun lexemes have the highest percentage, accounting for 60%, the 18082 verb lexemes about 37%, while the 1426 verb-noun lexemes for only 3%.

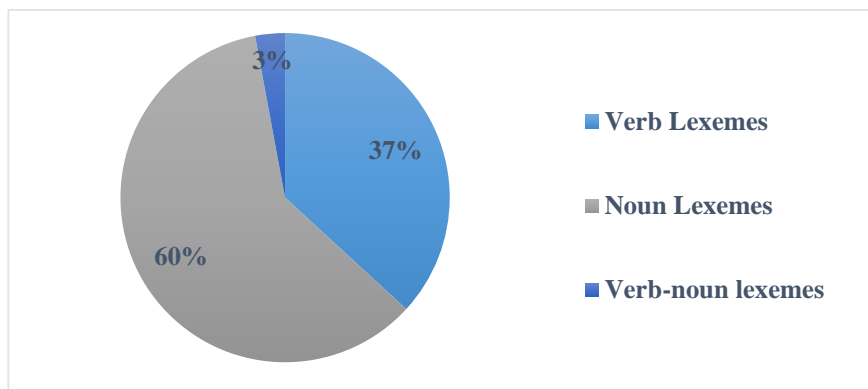


Figure 5. The percentage of verb lexemes, noun lexemes and verb-noun heterosemy in CCD7

Then, in order to compare with Shen's Model, a further step will be involved in the Model of Chinese verbs and nouns, which is presented bases on the usage pattern of verb lexemes, noun lexemes and verb-noun lexemes in CCD7 (Figure 7).

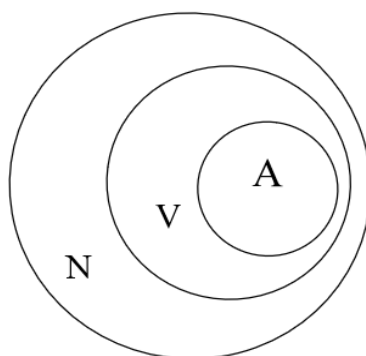


Figure 6. The Inclusive Model of Chinese content words (Shen, 2009)

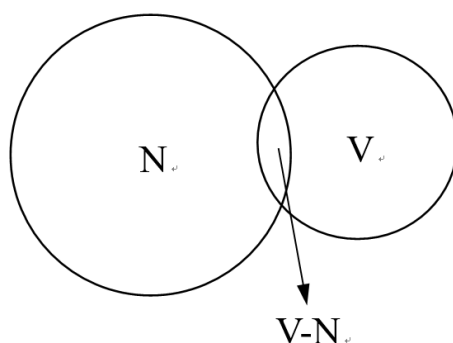


Figure 7. The Model of Chinese verbs and nouns

Figure 7 repeats the information from Figure 5, but it adds the boundaries of each individual word class. The Model of Chinese verbs and nouns here we propose (Figure 7) differs somewhat from the Model of Chinese content words (Figure 6) in terms of the dictionary-based pattern. By means of empirical testing, it is obvious that Shen's Model cannot hold water, it realizes that this model contradicts it. Firstly, the conclusions are well-supported by Figure 5, the Chinese word class system is not the inclusive model, not all verbs are included in nouns, words have their own

grammatical function in syntax and have fixed word classes in lexicon. Secondly, the wrong standard and procedure have been adopted to judge word class so that the most important point is neglected, that is, there is no clear-cut boundary between lexical items and syntagmatic patterns. Instead, the Model of Chinese verbs and nouns has not brought the problem about word classes to a grammatical system. This model seems to be self-evident and clearly indicates that nouns and verbs are different word class categorization, most of lexemes have fixed usage patterns, of which a few lexemes are labeled as nouns and verbs (verb-noun heterosemy). On the other hand, the important claim here is that even though the heterosemy is a minority, it is a “key minority”.

Using the analogy of quantum physics, Farrell (2001, p. 113-115) pointed out that N/V alternations (or functional shifts) is that words such as *bag*, *hammer*, *kiss*, and *sneeze* are basically neither nouns nor verbs, just as the physical entities (such as photon) cannot be appropriately characterized as either particles (things) or waves (processes); Rather, they manifest themselves as one or the other things depending on the context that is provided for their manifestation. Briefly, he held that words involved in N/V alternations have a relatively elaborate conceptual structure, and the class of a word is determined by the context of use; without the context, it has no class. Slightly different with Farrell, Martsa (2013, p. 56) claimed that these words are stored in the lexicon with their categories unspecified for verbhood or nounhood, while they receive full category specifications in or from the construction in which they are placed. In other word, Martsa realized that there are two levels in language, finally, these words with an unspecified form exist in lexicon. Dixon (2005, pp. 459-483), from the perspective of the form, meaning, adverb/adjective correspondence and preservation of peripheral constituents, took three main constructions as examples from the “HAVE/GIVE/TAKE A VERB”. Here, *have/give/take* substitute for the original verbs, which become (in base form of *VERB*) the head of a ‘second object’ NP, again preceded by the singular indefinite article *a*. He claimed that the words in the following of the construction “*have/give/take a*” express as verbs rather than the nouns.

However, Language is a complex adaptive system. Just as TLWCCT holds that word exists in two forms, that is, the word type in lexicon in communal language at *langue* and the word token in syntax at *parole*. With the repeated uses of word token, the class membership of the word type will be conventionalized. It then will be represented in lexicon with a fixed class (or form), and it is closely related to its frequency of use. It’s worth pointing out that the meanings of words in a dictionary are not the meanings at all. Rather, they are meaning potentials-potential contributions to the meanings of texts and conversations in which the words are used (Hanks, 2013, p. 73). Wang (2013, p. 12) also stated that multifunctionality of lexemes refers to the phenomenon that a polysemous lexeme has two or more word classes in the lexicon of the communal language system, which is represented as a heterosemous lexeme in language dictionaries; multifunctionality of lexemes is a common feature shared by analytic or isolating languages (Wang, 2014b, p. 55). On the other hand, language is the cultural heritage rather than a natural product, and language knowledge is derived from the constant use; the correlation between heterosemy and frequency is resulted from the results of economy and iconicity in communication, and the Principle of Parsimony is problematic in handling heterosemy of Chinese dictionaries (Wang & Zhou, 2015, p. 67). Relatively, this also explains well why these laws, such as the law of identity, the law of contradiction and the law of excluded middle, no longer take effect in the study of word class (see Wang & Yang, 2017, p. 32). The first order logic can only explain word class categorization in syntax, while the word class categorization at the communal language (in lexicon) requires high-order logic.

5.3 Case Study of the Word Class Labeling from the Two-level Word Class Categorization Theory

Language usage patterns influence language representation, language acquisition and language evolution (Beckner et al., 2009; Bybee, 2010). Obviously multicategory lexeme is a kind of language evolution caused by pattern changes and finally it is achieved by the language representation in the synchronic language level. Based on this, we further investigate the usage patterns of 斗争 ('struggle') in CCD7 (Figure 8) from the perspective of the Two-level Word Class Categorization Theory.

【斗争】 dòuzhēng ① 矛盾的双方互相冲突, 一方力求战胜另一方: 阶级~|思想~|跟歪风邪气做坚决的~。② 群众用说理、揭发、控诉等方式打击敌对分子或坏人: 开~会。③ 努力奋斗: 为祖国的繁荣昌盛而~。

Figure 8. Screenshot of the entry for 斗争 in CCD7 (v.)

A corpus-based case study is to be carried out. In CN corpus, 斗争 occurs 3183 times. Since the essence of word classes is the function of expression or propositional speech act (Croft, 1991; Guo, 2002), token frequency, type frequency, time span and register variation of lexemes are the main criteria to judge the word classes of lexemes (Wang & Chen, 2014). Thus, we mainly explore token frequency and type frequency of the word 斗争 respectively based on the overall sample. During the data processing, the use of word usage and word formation is firstly distinguished. In addition, of which 斗争性 ('fighting spirit'), 斗争会 ('public accusation meeting') and 斗争史 (lit, 'the history of fight'), 斗争 in here represents as word formation, and it occurs frequently and can be used independently. Thus, those cases are not considered in this study. The results are described in Table 6 and Table 7.

Table 6

The Token Frequency of 2598 Tokens 斗争 in CN Corpus

Propositional Speech Act	Word Usage		Morpheme Usage	Total
	Predication	Reference		
Number	381	2217	585	3183
Percentage	11.97%	69.65%	18.38%	100%

Table 7

The Type Frequency of 2217 Tokens 斗争 as Nouns in CN Corpus

Type	Number	Percentage	Examples
VP + 斗争	837	37.75%	揭露 /瓜分/争夺/争取/反对/加强 ~
NP + (的) + 斗争	777	35.05%	正义事业的/两条线路的/艰苦卓绝的 ~
Prep + 斗争	230	10.38%	在与/为了/经过 ~
斗争 + VP	36	1.62%	~ 是/变得/发展成/围绕
斗争(的) + NP	337	15.20%	~ 生活/ ~ 内容/ ~ 的需要/ ~ 的力量
Total	2217	100%	

Note. VP= verb or verb phrase; NP= noun or noun phrase; Prep=preposition.

According to the data in Table 6 and Table 7, the noun usage of 斗争 occurs 2217 times, accounting for 69.65%, while the verb usage repeats 381 times occupying 11.97%. Besides, the noun usage of 斗争 widely distributes in various structures, such as VP + 斗争, NP + (的) + 斗争, of which the percentage is 37.75% and 35.05% respectively. Evans and Green (2006, p. 118) pointed out that “frequency of use correlates with entrenchment; token frequency gives rise to the entrenchment of instances, type frequency gives rise to the entrenchment of more abstract schemas”. The results also confirm that token frequency is not the only criterion for judging the level of conventionalization (Wang & Chen, 2014, p. 26), type frequency is also an important determinant of productivity, with higher type frequency leads to greater productivity (Bybee, 2010, p. 95). 斗争, as a noun, bears highest token frequency and type frequency in CN Corpus. Specifically, it can be modified by various words and served as the object of many verbs. According to the Two-level Word Class Categorization Theory, we hold that the noun usage of 斗争 has conventionalized on the basis of its highest token frequency so that its noun usages are represented in various forms. Therefore, there is no doubt that the word 斗争 should be both labeled as a verb and a noun in the Chinese communal language system.

However, unfortunately, we can see that this is not case in Figure 8. CCD7 only label its verb usage, the noun usage is no admitted. In the same vein, 斗争 is marked as a verb in the CCD7, instead, the example, *跟歪风邪气做坚决的斗争* (lit, ‘(We) conduct a resolute struggle with the

unhealthy tendencies') shows that 斗争 is inappropriately marked as a verb. What has caused this phenomenon? Zhao (2015, pp. 118-119) elaborated that one of the reasons for conservative entry inclusion in the *CCD* lies in the corpora of Chinese were not sound, so that compilers had mainly depended on the compilers' own knowledge, as well as on some authoritative reference books, for the collection and selection of entries and entry examples. However, admittedly, even though compilers could accurately judge the typical members of a category, it is not easy to judge the atypical members. It might be more proper to say that a corpus can essentially tell us what language likes, intuition is a poor guide to judge a category (Hunston, 2002, p. 20). The essence of word class is the expressive function, and corpus-based investigation can grasp the facts of language and confirm whether it belongs to a word class. Therefore, the approaches for increasing the numbers of heterosemy and clarifying the nature of Chinese word class depend on the corpus-based usage model.

Finally, it proves by using the actual data that TLWCCT not only fills the gaps but also conforms to the rule of language evolution. According to TLWCCT, the lexemes in Chinese dictionary belong to the communal language, which label senses with word classes. In turn, the word token in the context belongs to the individual language, and it needs to be labeled according to the propositional speech act of the example in its particular use. The object of dictionary collects a conventional word, which represents the meaning potential of the individual language unit. The lexeme of the dictionary is usually regarded as a part of the communal language system.

6. Conclusion

This study has conducted the quantitative study of the representation of verb-noun heterosemy in *CCD7*. To sum up, there are both achievements and problems in the word class labeling of verb-noun heterosemy in *CCD7*. Generally, the number of verb-noun lexemes slightly increases. However, it still fails to fully reflect the status quo of verb-noun lexemes in Modern Chinese, and contradictions among word class labeling, the definition and the example are still prevalent. The study shows that, to a large extent, the bewilderment of verb-noun heterosemy in *CCD7* is a true portrayal of the dilemma of analytic languages, such as Modern Chinese. Scholars regard language as the natural object rather than the cultural inheritance, and they do not distinguish the relationship between the word type and the word token. Consequently, although *CCD7* is of high quality, it is undoubtedly influenced by the mainstream word class view of Chinese, especially in the processing of high frequency words. Heterosemy is the grammatical multifunctionality of lexemes in Modern Chinese, Modern English and other analytical languages, of which verb-noun heterosemy accounts for a large proportion and should thus be carefully considered. Similarly, the study of linguistics should learn from each other with the combination of contrastive linguistics and language typology. Modern Chinese dictionary should also learn more about the representation strategies of Modern English dictionary in order to continuously improve the quality of compilation and lay the foundation for teaching Chinese as an international language. As we know, the dictionary compilation is challenging and cumbersome, as such the flaws certainly cannot outweigh the virtues. This in turn suggests that more effects made to dictionary compilation are necessary and valuable.

Notes

① <http://www.aihanyu.org/cncorpus/index.aspx>

② The number of verb lexemes and noun lexemes in the study refers to the sum of the single category of lexemes and the two-categories lexemes (only in verb-noun lexemes) with labeling all senses with word classes.

③ The 49095 lexemes are the sum of verb lexemes, noun lexemes and verb-noun lexemes. The data here overlap with each other.

Acknowledgements

This study is supported by the Chongqing Social Sciences Foundation (No. 2014YBYY083) and the Chongqing Graduate Innovative Research Project (No. CYS18268). We hereby acknowledge the financial support of these projects.

References

A. Dictionaries

Chinese Social Sciences Academy. (2016). *The Contemporary Chinese Dictionary* (7th edition). Beijing: The Commercial Press.

Hornby, A. S. (Ed.). (2016). *Oxford Advanced Learner's Dictionary* (9th edition). Oxford: Oxford University Press.

B. Other literatures

Aarts, B. (2007). *Syntactic Gradience: The Nature of Grammatical Indeterminacy*. Oxford: Oxford University Press.

Baker, M. C. (2004). *Lexical Categories: Verbs, Nouns and Adjectives*. Cambridge: Cambridge University Press.

Beck, D. (2002). *The Typology of Parts of Speech Systems: The Markedness of Adjectives*. New York: Routledge.

Beckner, C. et al. (2009). Language is a complex adaptive system. *Language Learning* 59. s1: 1-26.

Bisang, W. (2011). Word classes. In J. J. Song (Ed.). *The Oxford Handbook of Language Typology* (pp. 280-302). Oxford: Oxford University Press.

Bloomfield, L. (1933). *Language*. New York: Holt.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Caramazza, A., & A. Hillis. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, 349 (6312), 788-790.

Chen, C. X., & Huang, J. H. (1993). Microstructure of bilingual dictionary (Part 1). *Modern Foreign Languages*, (4), 29-34.

Conwell, E., & J. L. Morgan. (2012). Is it a noun or is it a verb? Resolving the ambicategoricity problem. *Language Learning and Development*, 8 (2), 87-112.

Croft, W. (1991). *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: The University of Chicago Press .

- Croft, W., & E. van Lier. (2012). Language universals without universal categories. *Theoretical Linguistics*, 38 (1-2), 57-72.
- Crystal, D. (2010). *The Cambridge Encyclopedia of Language* (2nd ed.). Cambridge: Cambridge University Press.
- Dixon, R. M. W. (2005). *A Semantic Approach to English Grammar*. Oxford: Oxford University Press.
- Enfield, N. J. (2006). Heterosemy and the grammar-lexicon trade-off. In Ameka, F. K., Dench, A. C., & Evans, N. (Eds.). *Catching Language: The Standing Challenge of Grammar Writing* (pp. 297-320). Berlin: Walter de Gruyter.
- Enfield, N. J. (2015). *The Utility of Meaning: What Words Mean and Why*. Oxford: Oxford University Press.
- Evans, V., & M. Green. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh University Press.
- Fang, Q. M. (2010). Quantitative survey of cross-categorization of verb and noun in *Modern Chinese Dictionary*. *Lexicographical Studies*, (4), 30-40.
- Farrell, P. (2001). Functional shift as category underspecification. *English Language and Linguistics*, 5 (1), 109-130.
- Guo, R. (2002). *A Study of Chinese Word Classes*. Beijing: The Commercial Press.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.
- Harris, Z. S. (1946). From morpheme to utterance. *Language*, (3), 161-183.
- Herbst, T. (2010). *English Linguistics: A Coursebook for Students of English*. Berlin: Mouton de Gruyter.
- Heyvaert, L. (2003). *A Cognitive-Functional Approach to Nominalization in English*. Berlin/New York: Mouton de Gruyter.
- Hu, A. S. (2006). Delimitation between verbalized nouns and multifunctional words. *Language Teaching and Linguistic Studies*, (5), 1-6.
- Hu, M. Y. (1996). *Verb-Noun Words' Quantitative Research*. Beijing: Beijing Language and Culture University press.
- Hudson, R. A. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Ježek, E. (2016). *The Lexicon: An Introduction*. Oxford: Oxford University Press.
- Ježek, E., & Ramat, P. (2009). On parts-of-speech transcategorization. *Folia Linguistica*, 43 (2), 391-416.
- Jiang, Z. X., Ding, C. M., & Hou, Y. (2011). Bi-syllable words as both verbs and nouns in *Modern Chinese Dictionary* (5th ed.). *Lexicographical Studies*, (3), 37-50+71.
- Lefebvre, C. (2001). Multifunctionality and the concept of lexical entry. *Journal of Pidgin and Creole Language*, 16 (1), 107-145.

- Lieber, R. (2016). *English Nouns: The Ecology of Nominalization*. Cambridge: Cambridge University Press.
- Lichtenberk, F. (1991). Semantic change and heterosemy in grammaticalization. *Language*, 67 (3), 475-509.
- Liu, T., Ma, P. J., Yu, L., Liu, J. F., & Yang, Y. M. (2011). An event-related potential study of the noun-verb ambiguous effect in Chinese. *Journal of Psychological Science*, (3), 546-551.
- Lu, B. F. (1981). Issue on verbs-nouns lexemes: word class labeling of modern Chinese dictionary. *Lexicographical Studies*, (1), 151-155.
- Lu, B. F. (2012). The semantic characteristics of event nouns in Chinese and English. *Contemporary Linguistics*, 14 (1), 1-11.
- Lu, J. M. (2013). *A Course on Modern Chinese Grammar Research* (4th edition). Beijing: Peking University Press.
- Luuk, E. (2010). Nouns, verbs and flexibles: Implications for typologies of word classes. *Language Sciences*, 32 (3), 349-365.
- Lv, S. X. (1979). *Issues of Chinese Grammar Analysis*. Beijing: The Commercial Press.
- Ma, J. Z. (1898/1983). *Ma's Grammar*. Beijing: The Commercial Press.
- Malchukov, A. L. (2006). Constraining nominalization: function/form competition. *Linguistics*, 44 (5), 973-1009.
- Martsa, S. (2013). *Conversion in English: A Cognitive Semantic Approach*. Cambridge: Cambridge Scholars.
- Newmeyer, F. J. (1971). The source of derived nominals in English. *Language*, 47 (4), 186-796.
- Persson, G. (1986). Homonymy, polysemy and heterosemy: The types of lexical ambiguity in English. In Symposium on Lexicography III: Proceedings of the Third International Symposium on Lexicography, May 14-16, 1986.
- Quirk, R. et al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Rijkhoff, J. (2002). Verbs and nouns from a cross-linguistic perspective. *Rivista di Linguistica*, 14 (1), 115-147.
- Rijkhoff, J., & E. van Lier. (2013). *Flexible Word Classes: Typological Studies of Underspecified Parts of Speech*. Oxford: Oxford University Press.
- Robert, S. (2004). The challenge of polygrammaticalization for linguistic theory. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood, eds. *Linguistic diversity and language theories*. Amsterdam: Benjamins, 119-142.
- Robins, R. H. (1989). *General Linguistics: An Introductory Survey*. London: Longman.
- Sasse, H. J. (1993). Syntactic categories and subcategories. In J. Jacobs, A. von Stechow, W. Sternefeld, and T. Vennemann (Eds.). *Syntax: An International Handbook of Contemporary Research* (pp. 646-686). Berlin: Mouton de Gruyter.
- Shen, J. X. (2009). My view of word classes in Chinese. *Linguistic Sciences*, (1), 1-12.

- Shen, J. X. (2012). Reflections on “Nouny Verbs”: Problems and solutions. *Chinese Teaching in the World*, (1), 3-17.
- Shi, D. X. (2007). The criteria, method and result of the classification between noun and verb. *Chinese Language Learning*, (4), 3-12.
- Tan, J. C. (2018). Ideology of word-collection of language dictionary: Take examples for supplemented entries in *Contemporary Chinese Dictionary* (7th edition). *Studies of the Chinese Language*, (2), 223-225.
- van Lier, E. (2017). Lexical flexibility in Oceanic languages. *Studies in Language*, 41 (2), 241-254.
- Wang, L. (2016). Part-of-speech studies in Chinese. *Journal of Quantitative Linguistics*, 23 (3), 235-255.
- Wang, R. Q. (2006). *An Empirical Study of Word Class Labeling in Chinese-English Dictionaries from the Cognitive Perspective*. Shanghai: Shanghai Translation Publishing House.
- Wang, R. Q. (2010). A validity study of the word class system in Modern Chinese as seen from the *Contemporary Chinese Dictionary* (5th edition). *Foreign Language Teaching and Research*, (5), 380-386.
- Wang, R. Q. (2011). A cognitive study of wordhood in Modern Chinese based on *The Contemporary Chinese Dictionary* (5th edition). *Foreign Language and Literature*, (1), 71-77.
- Wang, R. Q. (2013). A study of multiple class membership in Modern Chinese with a comment on the significance of the linguistic theories of Ferdinand de Saussure. *Foreign Language and Literature*, (1), 12-20.
- Wang, R. Q. (2014a). Two-Level Word Class Categorization in analytic languages. In *Proceedings of the 36th Annual Conference of the German Linguistic Society, 2014*. Ed. University of Marburg, Germany, 345-347.
- Wang, R. Q. (2014b). Multiple class membership in Modern English: A study based on *Oxford Advanced Learner's Dictionary* (7th edition). *Journal of Foreign Languages*, (4), 50-59.
- Wang, R. Q., & Chen, H. M. (2014). A corpus-based study of the relationship between verbs and constructions: The conventionalization of transitive sneeze. *Foreign Language Teaching and Research*, (1), 19-31.
- Wang, R. Q., & Wang, D. (2016). A study of the representation strategy of verb-noun heterosemy in *Oxford Advanced Learner's English Dictionary* (8th edition). *Foreign Language and Literature*, (2), 51-58.
- Wang, R. Q., & Yang, X. (2017). The word class problem of chūbǎn (出版) and debates over endocentric constructions: A study from the perspective of the Two-level Word Class Categorization Theory. *Chinese Linguistics*, (4), 26-35.
- Wang R. Q., & Zhou, Y. (2015). A study of the correlation between heterosemy and frequency in Modern Chinese: A note on the validity of the Principle of Parsimony. *Foreign Language and Literature*, (2), 61-69.

Xu, S., & Tan, J. C. (2006). Remarks on the part-of-speech tagging in *The Contemporary Chinese Dictionary* (the 5th Ed.). *Chinese Language*, (1), 74-86.

Yang X. (2016). The overview on the study of word class labeling of heterosemy in *The Contemporary Chinese Dictionary* (the 5th and 6th editions). *Central China Normal University Journal of Postgraduates*, (3), 81-88.

Yang X. (2019). Strategies for the treatment of self-reference and partial self-reference in *A Dictionary of Contemporary Chinese* (7th edition). *Lexicographical Studies*, (1), 32-38.

Yang, X., & Wang, R. Q. (2018). Investigating the representation strategies of multi-category lexemes in *The Contemporary Chinese Dictionary* (6th edition). *Journal of Guangdong University of Foreign Studies*, (4), 5-13.

Yang, Y. M., Liang, D. D., Gu, J. X., Weng, X. C., & Feng, S. W. (2002). A neurolinguistic study on the classification of nouns and verbs in Chinese. *Linguistic Sciences*, (1), 31-46.

Yao, H. M. (1980). Talk about the boundaries between compound verbs and nouns. *Journal of Henan University*, (5), 81-86.

Yi, B. S., & Ni, C. B. (2018). Differentiation between nouns and verbs: Evidence from linguistics and neurocognitive science. *Journal of Tianjin Foreign Studies University*, 25 (6), 61-72.

Zawada, B. E. (2005). *Linguistic Creativity and Mental Representation with Reference to Intercategorical Polysemy*. Unpublished Doctoral dissertation, University of South Africa.

Zhang, Y. H., & Yong, H. M. (2007). *Modern Lexicography*. Beijing: The Commercial Press.

Zhao, G. (2015). The Contemporary Chinese Dictionary. *International Journal of Lexicography*, 28 (1), 107-123.

Zhu, D. X. (1982). *Lectures on Grammar*. Beijing: The Commercial Press.

THE TREATMENT OF PHRASEOLOGY IN CHINESE-ENGLISH DICTIONARIES: A PRELIMINARY STUDY

Zhang Xuhua

Fudan University

Abstract

There is little doubt that phraseology is at the heart of all language use. This paper examines the treatment of Chinese phraseology in two influential Chinese-English dictionaries. Two high frequency characters, 吃 and 打 (meaning “eat” and “hit” literally), were selected due to their highly polysemous and phraseological nature, and their phraseological behaviors examined in the Lancaster Corpus of Mandarin Chinese. The entries in the Chinese-English dictionaries for 吃 and 打 were examined and their content compared with the findings from the corpus-based study. The corpus-based identification and categorization of the phraseological behaviors of 吃 and 打 revealed that some multi-character expressions could not be covered by the terms offered by the existing taxonomy (Sag et al., 2002). Accordingly, the taxonomy was revised for the appropriate categorization of Chinese phraseology. The comparison of the corpus-based findings and entries showed that the overall lexicographical treatment of Chinese phraseology tends to be consistent in the two dictionaries. It was also found that the two dictionaries agree with each other on the overall inclusion and exclusion of phrases. It is worth noting that hardly any of the Verb-Particle Constructions observed in the corpus are included in the two dictionaries. We propose that these constructions should also be treated as phrases and it would be more user friendly if these phrases are not hidden in the other longer phrases, and are given the same status as the usual headwords. A larger corpus and sampling in the future would better characterize the taxonomy of Chinese phraseology and provide more conclusive findings.

Key Words: Chinese-English dictionary; phraseology; taxonomy; corpus

1. Introduction

Corpus research has now revealed a feature of language use which was grossly underestimated in previous descriptions, namely the pervasiveness of formulaic sequences, with their meanings and lexicogrammatical patterns. According to Erman and Warren (2000), more than half of both spoken and written texts were formulaic. Altenberg (1998) even claimed that the amount of formulaic sequences in native speakers’ speech to be more than 80%. The inclusion of a wealth of phrase information in the dictionaries has a huge impact on lexicography (Moon,

2007; Walker, 2009; Xu, 2013). It was argued that (Paquot, 2015) it is probably the lexicographical treatment of phraseology that corpus linguistics has had the most revolutionizing effect.

However, comparing with the developments in English phraseology and lexicography, the study of the lexicographical treatment of Chinese phraseology is still in its infancy (Xing, 2012; Li, 2014). Most studies in this field have only focused on the treatment of fixed multi-character expressions, including noun and verb phrases, idioms, proverbs etc. This indicates a need to examine the inclusion of semi-fixed multi-character expressions in the Chinese-English dictionaries.

In this paper, the focus of the observation is therefore put on the investigation of the lexicographical treatment of Chinese phraseology in two widely used Chinese-English dictionaries published in the past decade, namely, the 1st edition of the Chinese-English Dictionary (Unabridged) (Lu, 2015) (CEDLU) and the 3rd edition of the Chinese-English Dictionary (Wu, 2010) (CEDWU). Two high frequency characters, 吃 and 打 (meaning “to eat” and “to hit” literally), were selected and their phraseological behaviors examined in the Lancaster Corpus of Mandarin Chinese (LCMC), which provides a sound basis for monolingual investigations of Chinese (McEnery & Xiao, 2004). The lexicographical treatment of Chinese phraseology was observed by examining the entries in the two Chinese-English dictionaries, and then their content compared with the findings from the corpus-based study.

2. Data and Methods

2.1 High Frequency Characters

It is beyond doubt that frequency relates to several aspects of behavior and features of words and their meanings. As pointed out by Kilgarriff (1997: 135), a central fact about a word is how frequent it is. On the other hand, high frequency words are claimed to make the major contribution in text creation and be a major stumbling block for both lexicographers and learners of English due to their highly polysemous and phraseological nature (Sinclair, 1991; Nation, 2001; De Cock & Granger, 2004). Chinese characters, especially high frequency ones, are also notorious for their wide range of meanings and their productivity in word formation. Accordingly, 吃 and 打, which are among the top 30 most frequent verbs in LCMC, were selected and their phraseological behaviors examined in this study.

2.2 吃 and 打 and their Collocates in LCMC

According to McEnery and Xiao (2004:1175), LCMC, which is a one-million-word balanced corpus of written Mandarin Chinese, was designed as a Chinese match for the FLOB and Frown

corpora of British and American English. It contains five hundred 2,000-word samples of written Mandarin Chinese texts published in Mainland China around 1991.

Unlike English and other western languages, Chinese does not delimit words by space. Chinese words can be composed of multi-characters but with no space appearing between words. Word segmentation is therefore a crucial first step for Chinese processing tasks, such as wordlist generating and concordancing. Chinese word segmentation is quite complicated by the fact that there is no standard definition of word. Fortunately, LCMC is segmented and POS tagged, and therefore allows us to observe not only the phraseological behaviors of individual characters, but also the segmented words or chunks. It means the collocates between characters, words and even between characters and words could be analyzed in corpus LCMC.

Character	Segmented as characters and their collocates		Segmented as composing parts of words	
吃	478 times as characters	40 collocate types	吃饱蹲 吃力 吃饭	3 types and 123 tokens
打	410 times as characters	18 collocate types	打下 打破 打量 打开 打击 打断 打动	7 types and 224 tokens

Table 1 The frequency of 吃 and 打在 LCMC

吃 and 打 were segmented as single characters for 478 and 410 times and have 40 and 18 different collocation candidates respectively in LCMC (Table 1). The Mutual Information (MI) score was used as the prime statistics for filtering the collocation candidates from the WordSmith Tools generated collocates list. Collocates (including characters and words) co-occurring with a node-word in a span of 5:5 items to the left and the right of the node word were measured. Only characters or words with a minimum MI score of 3 were considered collocates of the node-word. Table 2 shows that the top-ten rank ordered collocates of 吃 and 打.

	吃			打		
	Collocate	Freq.	MI score	Collocate	Freq.	MI score
1	份饭	5	10.10	仗	5	9.87
2	早饭	7	9.89	招呼	12	9.83
3	饱	16	9.83	麻将	5	9.62
4	亏	9	9.63	桥牌	6	8.94
5	大锅饭	5	9.52	桌球	6	8.89
6	饭	36	9.31	主意	6	8.46
7	粥	6	9.20	电话	11	7.78
8	晚饭	6	9.04	定	6	6.10
9	苦	12	8.22	完	5	5.54
10	惊	8	8.0	水	6	4.70

Table 2 The collocates of 吃 and 打 (the top-ten rank ordered according to MI score)

It is very important to point out that the collocation candidates in Table 2 need to be further filtered because some items form transparent or compositional multi-character expressions with the node word 吃 or 打. By a compositional expression, we mean that the meaning of the expression is predictable by rule, as an intersection of the meanings of the constituent characters. In contrast, by a non-compositional expression, the meaning of the whole expression is more than the sum of its components: it has both internal grammatical structure and also semantic unity (Stubbs, 2001: 220). For example, 吃早饭 (to have breakfast) is a free combination of 吃 and 早饭 (breakfast) and is transparent in meaning, while 吃苦 (to endure hardships) (Figure 1) is semi-fixed and opaque in meaning.

、率先垂范、做好表率、<w POS="v">吃苦在前、享受在后；保持勤俭节约、艰苦朴

冷受不住，还能到广州<w POS="v">吃苦当兵，进讲武学校？！再说，我知道

，让小艾和孩子在家里<w POS="v">吃苦了，小艾受了感动，温温柔柔地哭了一

为顺姐是最勤劳、最肯<w POS="v">吃苦的人。重活儿、脏活儿她都干，每天

加强品德修养，培养乐于<w POS="v">吃苦、任劳任怨的高尚情操，牢固树立全心全意

和琉球人。琉球人很能<w POS="v">吃苦，这种非人的生活，他们也能捱得下去

终靠着这一群特别能够<w POS="v">吃苦的人，继续支撑了。霍英东在这个荒岛

总又说。“革命总是要<w POS="v">吃苦的，这个我懂，我不怕。”浦安修坚

Figure 1 A concordance of 吃苦 from LCMC

However, the borderline between compositional and non-compositional expressions is not always clear. Even an expression such as 吃饭 (to eat a meal) is sometimes semantically non-transparent. When it means “to make a living”, the propositional meanings of both 吃 and 饭 are weakened. Therefore, the division between compositional and non-compositional is somewhat subjective.

An observation of all the concordance lines shows that only 12 items form non-compositional multi-character expressions with 吃 (e.g., 吃饱, 吃苦 and 吃得) and 打 (e.g., 打仗, 打定 and 打水) respectively. Comparing with compositional expressions, non-compositional expressions, which are semantically opaque, present a bigger challenge to dictionary users and lexicographic practice. Accordingly, only non-compositional expressions were focused on in this study.

Furthermore, 吃 (3 types, 123 tokens) and 打 (7 types, 224 tokens) were sometimes segmented as composing parts of words (Table 1). For instance, 吃力 (laboursome) and 打开 (to open) are divided as words in text segmentation and therefore 力 and 开 are also taken as collocates of 吃 and 打. Although we have no statistical method to measure the collocation strength of the character pairs, it is clear from the segmentation that there is strong attractions between 吃 and 力, and 打 and 开.

To sum up, the concordance evidence shows that 14 and 19 items (including characters and words) occur frequently with 吃 and 打 to form non-compositional multi-character expressions (Table 3) which should be recorded in the Chinese-English dictionaries and will be examined in Section 3.

2.3 Research Procedures

The methodological steps generally shared in the previous studies (e.g., Moon, 2008; Walker, 2009), namely, description and comparison, were followed in the present study. The phraseological behaviors of 吃 and 打 were firstly examined in LCMC. Multi-character

expressions were identified by calculating the strength of collocates (MI scores) and scrutinizing the contexts of those collocates in section 2.2 above. Secondly, the entries of in the Chinese-English dictionaries for the two characters were examined with special attention on the documentation of their phraseological behaviors. Lastly, the observed content of 吃 and 打 in the dictionaries were then compared with the findings from the corpus-based study.

3. Results and Discussion

3.1 The Phraseological Patterns of 吃 and 打

Phrases or multiword expressions are named and categorized in different ways by researchers from different theoretical camps (e.g., Nattinger & DeCarrico, 1992; Lewis, 1993; Moon, 1998; Wray, 2005). According to Granger & Paquot (2008: 35), differences between the typologies largely correspond to differences in the selection of the features used to categorize multiword expressions and the prioritization of the selected features. One or more of the following features are given prominence in the classifications of the multiword expressions: (1) internal structure (e.g. verb + noun or verb + preposition); (2) extent: phrase- vs. sentence-level; (3) degree of semantic (non-)compositionality; (4) degree of syntactic flexibility and collocability; (5) discourse function. Consequently, various different typologies were developed to different purposes. Since this study is not intended as a comprehensive review of the different typologies of multiword expressions, we decided to use Sag et al. (2002)'s taxonomy of multiword expressions as a reference in the present study.

The compositionality of multiword expressions is crucial in Sag et al. (2002)'s taxonomy. Expressions were broadly categorized into lexicalized phrases, which have at least partially idiosyncratic syntax or pragmatics, and institutionalized phrases, which are syntactically and semantically compositional. As explained in the Section 2.2, only non-compositional expressions were selected and examined in this study.

The classification of multi-character expressions of 吃 and 打 are presented in Table 3 below. It is clear from the categorization that Sag et al. (2002)'s taxonomy was not followed closely. The structural characteristics of some multi-character expressions could not be covered by the terms offered by Sag et al. (2002). Accordingly, the Verb-Notional Constructions (including Verb-Noun, Verb-Adjective and verb-verb Constructions) was added to the taxonomy, while several categories (e.g., Compound Nominals and Proper Names) were excluded in our taxonomy because no expressions in those categories were found in this study.

	Categories and Types	Examples
Non-compositional Expressions (33)	Non-Decomposable Idioms (1)	吃大锅饭
	Decomposable Idioms (1)	吃饱蹲
	Light Verbs (2)	打招呼 打电话
	Verb-Notional Constructions (20)	吃亏 吃苦 吃东西 打水
		吃饱 打破 打开 打断 打动
		打量 打击
Verb-Particle Constructions (9)	吃得 吃点 (儿) 打起 打下	

Table 3 Classification of multi-character expressions of 吃 and 打

According to the parts of speech of the collocates of 吃 and 打 in the expressions, all the 33 non-compositional expressions, which are syntactically flexible, are classified into Non-Decomposable Idioms, Decomposable Idioms, Light Verbs, Verb-Notional Constructions and Verb-Particle Constructions (Table 3). The classification of Non-Decomposable Idioms and Decomposable Idioms was proposed by Sag et al. (2002). The difference between the two categories lies in if the process of semantic deconstruction starts off with the idiom and associates particular components of the overall meaning with its parts. For instance, 吃大锅饭 (literally to eat from the same big pot) is non-decomposable because it means “to mess” or “to get an equal reward or pay” in contexts. It does not associate with either 吃 (to eat) or 大锅饭 (collective dining) in a clear way. While 吃饱蹲 (literally be full and then to squat) is used to

refer to a person who is an idler or loafer and it has something to do with the literal meanings of 吃饱 (to have eaten one's fill) and 蹲 (to squat on the heels).

Light Verbs constructions normally consist of a verb (e.g., 进行 (to conduct), 做 (to do) and 给予 (to give)) and a noun. The verbs in the constructions carry little semantic content on their own. For instance, 打 (to hit) is semantically weak in 打招呼 (to greet, literally to hit greet) and 打电话 (to make a phone call, literally to hit telephone). By contrast, the verbs in the Verb-Notional Constructions are semantically stronger. They form predicates with additional expressions, which are usually nouns, adjectives and sometimes verbs. For instance, 吃 and 打 in 打仗 (to fight or to go to war)(Verb-Noun construction), 吃饱 (Verb-Adjective construction) and 打量 (to size up or to take the measure of)(Verb-Verb Construction) are followed by a noun (仗, warfare), an adjective (饱, full) and a verb (量, to measure). 吃 and 打 keep their semantic content to a certain extent in the Verb-Notional Constructions.

Verb-Particle Constructions, which are a subset of the verb-complements as delineated in Zhao (1979), are collocations of a verb and a particle. They are frequently followed by another element such as a noun or an adjective. For instance, 吃 and 打 co-occur with particles in 吃得 and 打下, and they frequently further collocate with adjectives (e.g., 好 (good), 兴高采烈 (in great delight), 痛苦 (pain or agony)) and nouns (e.g., 基础 (foundation)) respectively.

3.2 The Phraseological Treatment of 吃 and 打 by the Dictionaries

This preliminary study seeks to examine the lexicographical treatment of Chinese phraseology in two influential Chinese-English dictionaries. A comparison of the phrases included in the entries for 吃 and 打 shows that the lexicographical treatment of Chinese phraseology tends to be consistent in general in the two dictionaries. More than half (51.6% and 54.8% for CESLU and CEDWU respectively) of the phrases listed in Table 4 below appear in the dictionaries. This observation differs from those of the published studies (e.g., Moon, 2008; Walker, 2009) on the treatment of phraseology in the English dictionaries. For instance, it was revealed that no more than 20% of the total number of collocates or phrases listed for the selected items in the dictionaries appear in more than one dictionary (Walker 2009: 294). Apparently, there is a

lack of consistency in the selection of collocates by the dictionaries. This inconsistency between our result and that reported by Moon (2008) and Walker (2009) may be due to the different ways we deal with phraseological patterns in Chinese and English.

	CEDLU				CEDWU			
	吃		打		吃		打	
1	吃饱蹲	×	打破	√	吃饱蹲	×	打破	√
2	吃力	√	打量	√	吃力	√	打量	√
3	吃饱	×	打开	√	吃饱	√	打开	√
4	吃大锅饭	√	打击	√	吃大锅饭	√	打击	√
5	吃惊	√	打断	√	吃惊	√	打断	√
6	吃食	√	打动	√	吃食	√	打动	√
7	吃亏	√	打麻将	×	吃亏	√	打麻将	×
8	吃苦	√	打桌球	×	吃苦	√	打桌球	×
9	吃东西	×	打桥牌	×	吃东西	×	打桥牌	√
10	吃不	×	打仗	√	吃不	×	打仗	√
11	吃得	×	打招呼	√	吃得	×	打招呼	√

12	吃了	×	打主意	√	吃了	×	打主意	√
13	吃点(儿)	×	打定	√	吃点(儿)	×	打定	×
14	吃下去	×	打电话	√	吃下去	×	打电话	√
15			打水	√			打水	×
16			打起	×			打起	×
17			打个	×			打个	×
18			打得	×			打得	×
19			打下	×			打下	√

(√ = included as a phrase; × = not included)

Table 4 The inclusion of expressions made up of 吃 and 打 and their collocates in the dictionaries

Another consistency is that in most cases (87.1%, 27 out of 31) the two dictionaries agree with each other on the overall inclusion and exclusion of phrases. It means that the lexicographers only made four different decisions on the treatment of the phraseological patterns of 吃 and 打. Among the few exceptions are 吃饱, 打定 (to decide on) and 打水 (to draw water) from the Verb-Notional Constructions category.

There is also a noticeable similarity in the way in which the Verb-Particle Constructions are excluded in the two dictionaries. It is clear from Table 4 that almost all (94.5%, 17 out of 18) the Verb-Particle Constructions are not included in the dictionaries. Verb-Particle

Constructions are not taken traditionally as well-established phrases in most instances. The analysis in section 3.1 shows that the Verb-Particle Constructions co-occur frequently with adjectives and nouns. Without these notional words, the Verb-Particle Constructions are usually not assumed to function as semantic units. They fail an important and widely accepted criterion used to define phraseological units. For instance, Gries (2008: 4) proposed six parameters in defining phraseology and one of them being ‘semantic unity’, i.e. to have no sense like a single morpheme or word. It is clear from the concordance lines that most Verb-Particle Constructions do not form strong semantic units and have no sense like a single character or word.

On the other hand, the Verb-Particle Constructions meet almost all the other criteria made, e.g., frequency of occurrence, lexical flexibility and non-compositionality. For instance, 吃点 (儿) (儿 is a noun suffix carrying a colloquial tone) can be analysed with respect to the five out of six criteria proposed by Gries (2008: 4):

Nature of the elements: characters;

Number of elements: two

Frequency of occurrence: the two parts of the expression co-occur more often than expected by chance;

Distance of elements: the two parts of the phraseologism usually co-occur adjacently (in 3 out of 8 cases, where 了 intervenes);

Flexibility of the elements: 了 is intervened to change the expression from the present to past tense, but 点 can never be preposed.

The concordance evidence (Figure 2) shows that although 吃点 (儿) could not be replaced by a single character or word, it resembles the English transitive phrasal verbs (Gries, 2008: 7), e.g., to give up and to give up, especially in terms lexical and syntactic flexibility and non-compositionality. It roughly means ‘to eat or take’ and ‘to suffer’ in different contexts. It collocates with nouns such as 东西 (thing or stuff), 什么 (something), 干粮 (ready-to-eat pre-packaged food) and 腐乳 (fermented bean curd) to mean “to eat something”, while with nouns such as 苦 (hardship) and 亏 (deficit) to mean “to have a rough time” and “to suffer a loss”.

Most importantly, 点 (儿) is used to emphasize the small amount of things to eat or limited

degree of hardship to endure. The structure and meaning of the phrase would not be changed significantly if it is dropped.

"阿兰，等一会儿还是先吃点东西，好吗？"项青项兰都注意

厚实的胸脯："没问题。吃点苦、受点累，我这身体还行。"

是饱汉子不知饿汉子饥。吃点苦算什么？我是不甘心呀，说来说

了，两位仙长吃点亏就吃点亏吧，俗话说，'不打不相识'嘛

云彩都散了，两位仙长吃点亏就吃点亏吧，俗话说，'不打

叫一声阿毛早上好，想吃什么呢？啊，要一份净面外卖，

失手就会全盘败北，轻者吃点小亏；重者债台高筑，甚至自杀身亡

商旅，到驿站打个尖儿，吃点自备的干粮，喝些驿卒自酿的水酒 骂他没本事，老婆怀孕想吃点儿臭腐乳都吃不到。

睡到凌晨，李

Figure 2 A concordance of 吃点 (儿) from LCMC

Accordingly, multi-character expressions such as 吃点 (儿) and 打起 in the Verb-Particle Constructions category are regarded as phrases in this study and it is proposed that they should be included as subentries of 吃 and 打 in the Chinese-English dictionaries.

However, Verb-Particle Constructions are usually treated as composing parts of longer phrases in both CEDLU and CEDWU. For instance, both CEDLU and CEDWU have treated 吃得 within subentries for the verb 吃, as part of longer phrases, such as 吃得开 (to be popular), 吃得来 (to be able to eat), 吃得上 (to afford to eat), 吃得下 (to be able to eat), 吃得消 (to be able to endure), 吃得住 (to be able to support or bear) and 吃得准 (to figure out or calculate accurately). Since most of the adjective characters collocate with 吃得 to form phrases with a positive meaning, it would be helpful to treat 吃得 as a subentry under the verb 吃 and list the collocates that co-occur frequently with it in the corpus, e.g., 开, 来, 上, 下, 消, 住 and 准.

4. Conclusion

The aim of the study was to investigate the lexicographical treatment of Chinese phraseology in two widely used Chinese-English dictionaries (CEDLU and CEDWU). The focus of the observation was put on the investigation of the phraseological behavior of high frequency characters included in the dictionaries. Two high frequency characters, 吃 and 打, are firstly examined in LCMC to identify their phraseological behaviors. The phrases found in the corpus-based studied were then compared with the content in the dictionaries.

It was found that the lexicographical treatment of Chinese phraseology tends to be consistent in general in the two dictionaries. Another consistency is that the two dictionaries agree with each other on the overall inclusion and exclusion of phrases. It is worth noting that almost all the Verb-Particle Constructions are excluded from the two dictionaries. We propose that these constructions should also be treated as phrases and it would be more user friendly if these phrases are not hidden in the other longer phrases, and are given the same status as the usual headwords, as it was initiated by Sinclair (2010).

Furthermore, it was revealed that the identification and categorization of Chinese phrases should be based on the examination of corpus evidence. Sag et al. (2002) 's taxonomy of multiword expressions was revised with categories added and deleted. However, the current study is limited by the size of the corpus used and studies based on larger corpora would more

definitively characterize the taxonomy of Chinese phraseology. A further limitation is that only two high frequency characters were selected and examined, whereas a larger sampling would provide more conclusive findings.

References

Dictionaries

1. Lu G (陆谷孙). (2015). *中华汉英大词典(The Chinese-English Dictionary-Unabridged)*. 复旦大学出版社.
2. Wu G (吴光华). (2010). *汉英大辞典(The Chinese-English Dictionary)*. 上海交通大学出版社.

Other Literature

1. Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. *Phraseology: Theory, Analysis and Applications*. Oxford: OUP, 101-124.
2. De Cock, S., & S. Granger. (2004). High Frequency Words: the Bête Noire of Lexicographers and Learners A Like.
3. Erman, B. & B. Warren (2000). The idiom principle and the open choice principle. *Text* 20(1): 29-62.
4. Granger, S., & M. Paquot. (2008). Disentangling the phraseological web. *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins, 27-49.
5. Gries, T. (2008). Phraseology and Linguistic Theory: A Brief Survey. *Phraseology: An interdisciplinary perspective*. Amsterdam: Benjamins, 3-25.
6. Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135-155.
7. Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Hove: Language Teaching Publications.
8. Li, D. (李德俊). (2014). 短语及其自动识别研究评述. *外语研究*(6), 8-13.
9. McEnery, A. M., & R. Z. Xiao. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *language resources and evaluation*.
10. Moon, R. (2007). Sinclair, lexicography, and the Cobuild Project: The application of theory. *International journal of corpus linguistics*, 12(2), 159-181.
11. Moon, R. (2008). Dictionaries and collocation. *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins, 313-336.

12. Nation, I. (2001). How many high frequency words are there in English. *Language, learning and literature: Studies presented to Hakan Ringbom. Abo Akademi University, Abo: English Department Publications, 4*, 167-181.
13. Nattinger, J. R. & J. S. DeCarrico (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
14. Paquot, M. (2015). Lexicography and phraseology. *The Cambridge Handbook of Corpus Linguistics*. Cambridge University Press: Cambridge. <http://hdl.handle.net/2078.1/139795>.(accessed 10 January 2018) Google Scholar.
15. Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *international conference on computational linguistics*, 1-15.
16. Sinclair, J. (1991). *Corpus, concordance, collocation*: Oxford University Press.
17. Sinclair, J. (2010) Defining the definiendum. *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Kampala and Ghent: Menha Publishers: 37-47.
18. Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. New York: Wiley-Blackwell.
19. Walker, C. (2009). The treatment of collocation by learners' dictionaries, collocational dictionaries and dictionaries of business English. *International Journal of lexicography*, 22(3), 281-299.
20. Wray, A. (2005). *Formulaic language and the lexicon*: Cambridge University Press.
21. Xu, H. (徐海). (2013). 短语学与英语短语的词典收录及编排. *辞书研究*, (5), 46-52.
22. Xing, F. (邢富坤). (2012). 多词单位的描写识别与词典编纂. *当代语言学*(4), 407-417.
23. Zhao, Y. (赵元任). (1979). *汉语口语语法*. 商务印书馆.

LMF RELOADED

Laurent Romary^{1,3,4}, Mohamed Khemakhem^{1,2,3}, Fahad Khan⁸, Jack Bowers^{1,6,7}, Nicoletta Calzolari⁸, Monte George⁵, Mandy Pet⁵, Piotr Bański⁹

1. Inria-ALMAnaCH - Automatic Language Modelling and ANalysis & Computational Humanities
2. UPD7 - Université Paris Diderot - Paris 7
3. CMB - Centre Marc Bloch
4. BBAW - Berlin-Brandenburg Academy of Sciences and Humanities
5. ANSI- American National Standards Institute
6. EPHE - École Pratique des Hautes Études
7. ÖAW - Austrian Academy of Sciences
8. CNR-ILC - Istituto di Linguistica Computazionale "Antonio Zampolli"
9. IDS - Institut für Deutsche Sprache

Abstract

The Lexical Markup Framework (LMF) or ISO 24613 [1] is a *de jure* standard which constitutes a framework for modelling and encoding lexical information both in retrodigitised print dictionaries as well as in NLP lexical databases. An in-depth review is currently underway within the standardisation sub-committee, ISO-TC37/SC4/WG4 with the goal of creating a more modular, flexible and durable follow up to the original LMF standard published by ISO in 2008. In this paper we will showcase some of the major improvements which have so far been implemented in the new version of LMF.

Key Words: ISO 24613, LMF, Lexical resources

1. Introduction

The previous version of LMF, published by ISO in 2008 [1] offered a framework for modelling, publishing and sharing lexical resources with a special focus on requirements arising from the domain of Natural Language Processing (NLP). Due to the potential richness and the multi-layered nature of linguistic descriptions in lexical resources the LMF meta-model ended up taking on a great deal of complexity in its attempt to reflect these various different linguistic facets. At the same time key areas of linguistics such as etymology (and diachronic lexical information in general) were not covered at all. Finally, the recommended serialisation for LMF was not clearly compatible with other leading markup standards, namely the TEI [3].

For these, and other reasons it was decided that the standardisation sub-committee, ISO-TC37/SC4/WG4 should review the LMF meta-model in order to create a new version of the standard which would address all of these issues. This new version of LMF will constitute a multi-part standard consisting of seven modules with the possibility of further extensions. Importantly the new version of LMF will be backwards compatible with the 2008 version. In the following section we will describe each package of the revised standard.

2. Abstract Modelling

In keeping with the fundamental conceptual modelling principles which have been decided on by ISO-TC37/SC4/WG4, the proposed model has been decoupled from any single serialisation format, although two potential serialisations of the meta-model constitute parts iv and v of the standard (TEI and LBX respectively). As a result, three major improvements have been carried out: restructuring; enrichment and simplification. Each are discussed below.

Restructuring

Although the previous version of the standard reflected some separation among packages touching on different linguistic levels of description, the differentiation lacked a sufficient level of modularity. A user of the standard had to get the standard as a whole package where he could be interested in just specific parts of it. The current version of the standard is much more modular and has been split into the following seven parts:

- i. ISO 24613-1 - Core model: defines basic classes required to model a baseline lexicon
- ii. ISO 24613-2 - Machine Readable Dictionaries (MRD) model: contains components providing deeper specification of lexical description encapsulated within the core model. *Form* is for instance differentiated into *Related Form*, *Word Form*, *Stem* and *Word Part*
- iii. ISO 24613-3 - Diachrony-Etymology: categories related to word and meaning origin and change are defined
- iv. ISO 24613-4 - TEI serialisation: represents a first serialisation of the first three parts based on a restricted version of the Text Encoding Initiative (TEI) guidelines [2]
- v. ISO 24613-5 - LBX serialisation: a second serialisation is formalised here using Language Base Exchange (LBX)
- vi. ISO 24613-6 - Syntax and Semantics: semantic and syntactic components are gathered in this extension to be revised and integrated with the first three parts of the standard
- vii. ISO 24613-7 - Morphology: morphology package will be defined in a separate part of the standard and will also be interconnected with the first three parts of the standard

The restructuring comes along with a revision of class membership; for example, the *Lemma* class, which was previously based in the MRD part is now part of the Core Module as it is a fundamentally essential part of a lexicon.

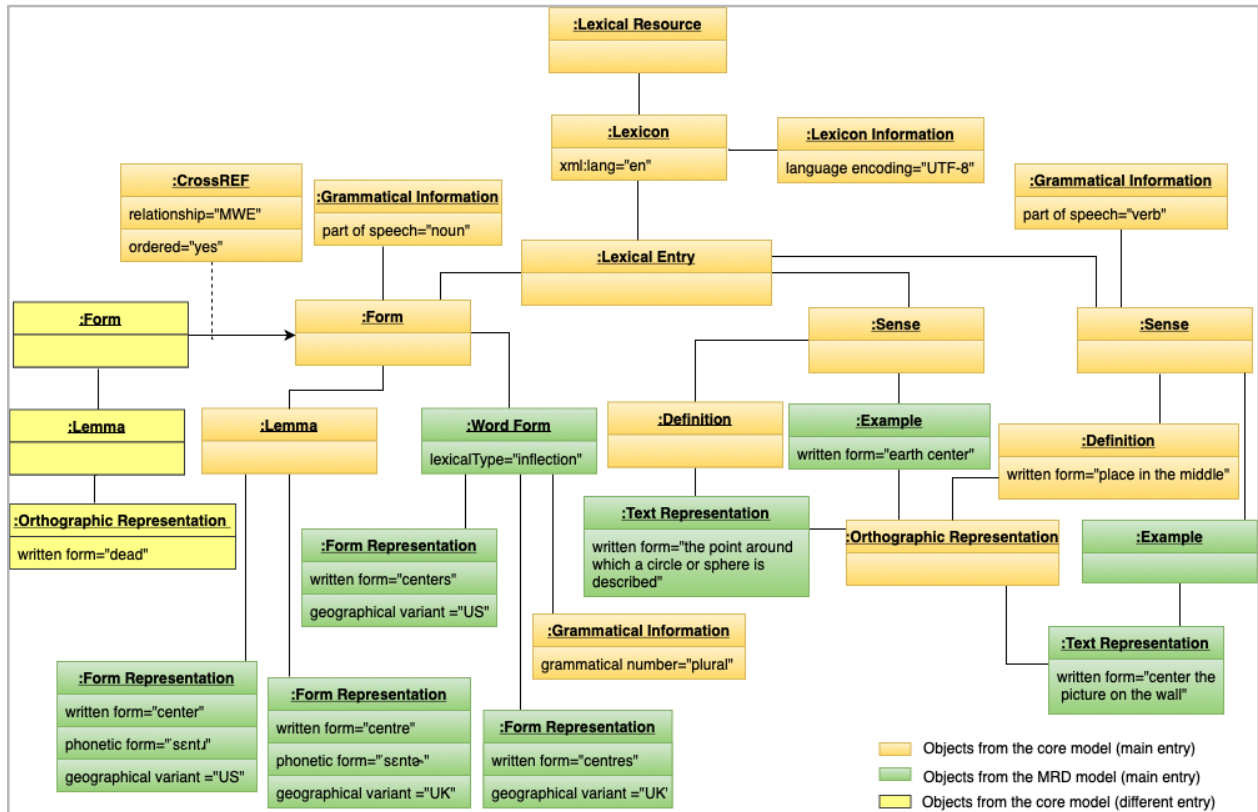


Figure 1: Example of the lexical entry “Center” encoded using the core (ISO 24613-1) and MRD (ISO 24613-2) metamodels

b. Enrichment

New information has been introduced to describe essential aspects of lexical information such as *Bibliography*. Such information is required to specify references for some usages, definitions, examples, etc. Therefore the new class is kept multi-functional to be used in case of need as determined by the editor of a lexicon. Additionally, the differentiation of Orthographic Representation into Form Representation and Text Representation has been designed to enable more precision in the encoding of written forms touching respectively Sense and Form sub-classes.

c. Simplification

The emphasis on abstraction and modularisation has also led to a series of major simplifications affecting nearly every class of the meta-model. One key feature which is being newly introduced is the CrossREF class which is a pointing/mapping mechanism that can be used to model a wide array of lexical features and relationships such as semantic relations, cross references, related entries and others within the meta-model. As a result some classes (e.g. *List of Components* and *Component*) whose features have been taken on, in part, by *CrossREF* have been removed altogether. Figure 1 illustrates the simplicity of the new

mechanism used to model Multi Word Expressions (MWE) previously represented by classes which are now obsolete.

3. More Coverage

One of the main intentions behind the new version of the standard is to provide increased coverage of the type of information that can be encoded by the model. To this end a completely new meta-model covering etymological and diachronic information is proposed in (ISO 24613-3). This new module extends the core LMF and MRD metamodels adding the following key classes (among others):

- **Etymon** and **Cognate**: subclasses of the core class Lexical Entry which are used in describing the diachrony of other lexical entries. Etymons are lexical entries from which another lexical entry is derived (a historical form or sense), and Cognates are lexical entries in related languages which share a common ancestor with a given aspect of a lexical entry.
- **EtyLink**: subclass of CrossREF which is used to link one or more temporal stages of one or more aspects of a lexical entry (i.e. sense, phonetic properties, etc.)
- **Etymology**: describes the history of a lexical entry or other element by being associated with an ordered series of EtyLink instances. An Etymology instance can be recursive and typed to define the changes undergone according to any number of linguistic processes (e.g. borrowing, inheritance, metaphor, metonymy, etc.)
- **Date**: defines specific or relative temporal information associated with some aspect of an etymology or its components

We illustrate the new etymology module with an example, an etymology for the word *center* which we have taken from Klein [4]:

center, centre, n. — F. *centre*, fr. L. *centrum*, fr. Gk. κέντρον, 'point, prickle, spike, ox goad, point round which a circle is described', from the stem of κέντειν, 'to prick, goad', whence also κέντωρ, 'a goader, driver', κεστός (for *κεντρός), 'embroidered', κέστρα, 'pickaxe', κοντός, 'pole', fr. I.-E. base **kent-*, 'to prick', whence also Bret. *kentr*, OIr. *cinteir*, 'a spur', OHG. *hantag*, 'sharp, pointed', Lett. *sīts*, 'hunter's spear', *situ, sist*, 'to strike', W. *cethr*, 'nail'. Cp. **centrifugal, centripetal, concentrate, eccentric, Dicentra, paracentesis**. Cp. also **cestrum, cestus**, 'girdle', **kent**, 'a pole', **quant**, 'a pole'.
Derivatives: *center, centre*, intr. and tr. v., *center-ing, centr-ing, centre-ing*, n.

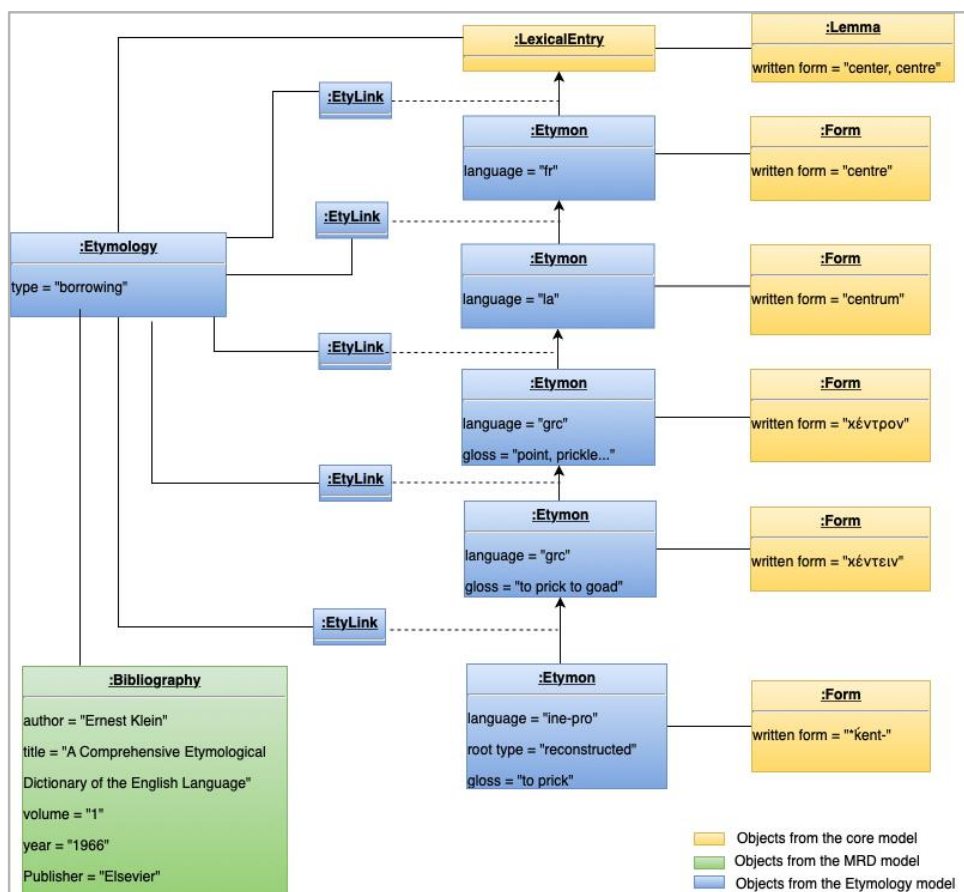


Figure 2: Example of modelling Etymological information following ISO 24613-3 meta-model, together with the entry being modelled from [4]

In Figure 2 we present an LMF model of a part of the Klein entry, namely, the portion that describes the borrowing of the word *center* from French, and its history prior to that. Note that for reasons of space we haven't shown the ordering of the Etylinks in the diagram nor specified the types of link (specifying for instance if an etymological link between two elements represents an etymological borrowing or inheritance). This is however an important aspect of etymological encoding and should be included in LMF resources describing the etymologies of lexical entries.

4. Connection with Leading Standards

The Text Encoding Initiative (TEI) is a standard which is widely adopted among lexicographers. The new version of LMF contains a TEI serialisation which aims to make both standards fully compatible following the vision presented by Romary in [3]. This serialization has the benefit of being able to leverage the knowledge and make use of the established practices of the TEI community in dealing with the representation of a wide array of lexicographic issues with which LMF is also concerned.


```

<entry>
  <form type="lemma" xml:id="center_form">
    <orth>center</orth>
    <pron>'sentʃ</pron>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
    <usg type="geo">U.S</usg>
    <form type="variant">
      <orth>centre</orth>
      <usg type="geo">U.K</usg>
      <pron>'sentə</pron>
    </form>
  </form>
  <form type="inflected">
    <orth>centers</orth>
    <usg type="geo">U.S</usg>
    <gramGrp>
      <number>plural</number>
    </gramGrp>
  </form>
  <form type="inflected">
    <orth>centres</orth>
    <usg type="geo">U.K</usg>
    <gram type="number">plural</gram>
  </form>
  <sense>
    <def>the point around which a circle or sphere is described</def>
    <cit type="example">
      <quote>earth center</quote>
    </cit>
  </sense>
  <sense>
    <gramGrp>
      <pos>verb</pos>
    </gramGrp>
    <def>place in the middle</def>
    <cit type="example">
      <quote>center the picture on the wall</quote>
    </cit>
  </sense>
  <re type="multiWordExpression">
    <form>
      <seg corresp="#dead_form" n="1">dead</seg>
      <seg corresp="#center_form" n="2">center</seg>
    </form>
  </re>
</entry>

```

Figure 3: Encoding example following LMF's TEI serialisation (ISO 24613-4)

The TEI guidelines offer a great degree of freedom for encoding lexical information. However in some cases, such freedom comes at the cost of an excess of variability in how users choose to represent certain features. Therefore in this serialization, we have sought to constrain that flexibility in line with similar initiatives, namely TEI Lex-0 [5,6,7]. In Figure 3 we show how the components in Figure 1 can be serialised using TEI elements. The development of a list⁷⁴, gathering serialisation examples provided by the community and checked by the ISO experts, along with a schema specification⁷⁵ is underway.

5. Conclusion

In this work we have presented the measures we followed to remedy the deficiencies noted in LMF after years of release. The changes, being in some cases important, will bring more flexibility and interoperability to the standard. The in-depth structure review along with the enriching new modules will be great assets for the current users of the standards and hesitant lexicographers to adopt the standard as they could benefit not only from the advantages of a de jure standard like ISO 24613 but also to have efficient modelling and serialisation alternatives.

References

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation* 43, no. 1 (2009): 57-70. .

TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. 15.03.2019]. <http://www.tei-c.org/P5/>.

Laurent Romary. TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics*, 2015, 30

Ernest Klein. *A Comprehensive Etymological Dictionary of the English Language*. Volume 1. 1966. Elsevier.

Laurent Romary and Toma Tasovac T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. *JADH 2018*. 2018 Sep 9:274.

Piotr Bański, Jack Bowers, Tomaz Erjavec. TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, Sep 2017, Leiden, Netherlands.

Jack Bowers, Laurent Romary. Encoding Mixtepec-Mixtec Etymology in TEI. *TEI Conference and Members' Meeting*, Sep 2018, Tokyo, Japan.

⁷⁴<https://github.com/DARIAH-ERIC/lexicalresources/blob/master/Schemas/LMFinTEI%20Specification/examplesLMFinTEI.xml>

⁷⁵<https://github.com/DARIAH-ERIC/lexicalresources/blob/master/Schemas/LMFinTEI%20Specification/LMFinTEIspec.html>

FROM CONCEPT DEFINITIONS TO SEMANTIC ROLE LABELING IN SPECIALIZED KNOWLEDGE RESOURCES

Ivana Brač and Ana Ostroški Anić

Institute of Croatian Language and Linguistics

Abstract

The paper presents the framework for semantic role labeling that is being developed within the project the *Dynamicity of Specialized Knowledge Categories*. The main goal of the project is the description of the categories of aviation within semantic frames. One of the tasks towards achieving this goal is developing a methodology for the syntactic and semantic analysis of the language for specialized purposes that could be applied to any specialized domain. The paper first outlines the differences between the terminological database Struna and a lexical database of semantic frames of aviation, AirFrame. Methods applied in analyzing terminological definitions and annotated sentences containing aviation terms are then presented. Conceptual information from the definitions of key aviation concepts in Struna is compared to semantically annotated sentences extracted from the parallel English-Croatian aviation corpus. The corpus compilation and analysis has been done in Sketch Engine, while the annotation is done in the WebAnno tool. Two approaches to semantic role labelling are discussed. The first approach does not have roles specific to one class, e.g. LIRICS (Petukhova & Bunt, 2008) and VerbNet (Kipper et al., 2008) with hierarchically organized semantic roles. The second approach, developed within FrameNet has more than 2000 frame elements or semantic roles that are verb-specific and frame-specific. Advantages of each approach are discussed, and the semantic tagset applied in the analysis is elaborated. The benefits of using semantic role labeling in terminological resources are finally discussed.

Keywords: terminology, semantic roles, semantic role labeling, specialized knowledge.

1. Introduction⁷⁶

The development of specialized knowledge resources relies heavily on using the existing tools for general language description, and on integrating good lexicographic practice and theoretical findings in such a way so as to establish a link between general and specialized knowledge. Online terminological dictionaries and databases that exploit the semantic-syntactic potential of terminological units, thus trying

⁷⁶ This work has been fully supported by the Croatian Science Foundation (www.hrzz.hr) under the project UIP-2017-05-7169.

to reflect the nature of specialized knowledge categories, are becoming more and more recognized (Faber, 2012; L'Homme, 2012,). Often is the work of developing a specialized database of a particular domain the continuation of previous, more traditionally oriented terminology work, as is the case of AirFrame, a database of semantic frames in the field of aviation that is being created within the research project the *Dynamicity of Specialized Knowledge Categories* (DIKA). One of the tasks towards achieving this goal is developing a methodology for the syntactic and semantic analysis of the language for specialized purposes that could be applied to any specialized domain. In order to compile a set of semantic roles applied in the process, a comparison between terminological definitions of aviation concepts in Struna and sentences extracted from a specialized corpus has been made.

Struna is a terminological database developed within the program the *Development of the Croatian Special Field Terminology* (known under its Croatian acronym Struna), financed by the Croatian Science Foundation (HRZZ). The program started in 2007, and is being carried out at the Institute of Croatian Language and Linguistics. The main purpose of Struna is to standardize Croatian terminology of various professional domains, and make it available to the public through a national term bank developed in-house for this purpose (Brač, Bratanić, & Ostroški Anić, 2015). As most normative term banks the primary task of which is defining and prescribing terminology in languages with a less standardized terminological component, Struna is largely based on the traditional terminological premises set out in the work of Eugen Wüster and his followers. According to these terminological principles, codified in the ISO terminology standards, the terminology of a certain special field is defined as a structured group of concepts, concept relations and terms as their designations (ISO 704). However, although the model of the terminology standardization applied in Struna originally rather strictly adhered to the semiotic principles defined by this terminological tradition, certain accommodations have been made over the course of years, both in data categories and the methodology applied, that moved Struna more in the direction of a descriptive end of terminology work.

The organization of the terminology workflow required a proper terminology management system; therefore an in-house solution was designed in order to cover both the need for an editing and storage tool, as well as for a search and retrieval application. In order to make Struna compatible and exchangeable with the existing termbases, its structure was designed in accordance with the TEI P5 guidelines for text mark-up and the TBX standard format for the representation and exchange of terminological data (Melby, 2015). The current list of record elements includes the following data categories: subject field and subfield, preferred Croatian term, source of the term, foreign term language label, neologism label, interdisciplinary term label, grammatical information on the preferred term, definition and its source, context with its source, synonyms according to their normative status (admitted, proposed, deprecated, obsolete, colloquial), equivalents in other languages, subordinate concept, abbreviation in Croatian and other languages, symbol, formula, equation, hyperlink, picture, note, and a field for correspondence among the domain editors, terminologists, and language experts (Bratanić & Ostroški Anić, 2013).

Unlike Struna, which is a multidomain terminological database, AirFrame is a monodomain terminological resource. It is a lexical database with a terminological function in which aviation concepts and their definitions, terms in several languages, and other relevant categories of terminological entries are presented in specialized semantic frames. Such a presentation of specialized knowledge is dynamic because

it is based on situationally, contextually and culturally conditioned semantic frames (Fillmore, 1982; Fillmore, Johnson, & Petruck, 2003). The database will consist of these categories: semantic frames, frame elements with their definitions, terms and terminological units such as collocations and phraseological units (corresponding to lexical units in FrameNet), sentences labeled with semantic roles, and frame-to-frame conceptual relations. Following the general methodology of FrameNet (Ruppenhofer et al., 2010), semantic frames and their elements are defined separately as opposed to lexical data.

2. Methods

Definitions of key aviation concepts as defined in Struna were analyzed and compared to sentences containing terms for those concepts, in order to establish the difference in conceptual and semantic information that can be gathered from terminological definitions as opposed to annotated examples from corpora. The list of key aviation concepts used as target words for extracting sentences included: *aerodrome, aircraft, airline, airspace, air traffic, air traffic control, air transport, flight, landing and taking off*. Definitions were analyzed for concept characteristics, and annotated for ontological categories contained in them, which make frame elements. A parallel English-Croatian aviation corpus compiled within the project was then queried for English and Croatian terms of the analyzed aviation concepts. Extracted sentences were annotated for semantic roles using the LIRICS semantic tagset (Petukhova & Bunt, 2008). Corpus analysis was done in Sketch Engine (Kilgariff et al., 2014), while the annotation was carried out using the WebAnno tool (Eckart de Castilho et al., 2016).

3. Results

Since traditional terminology work is largely focused on defining entities as prototypical categories for knowledge representation, it was expected that entities would comprise the largest share of the types of categories labeled as frame elements in definitions. Apart from entities such as *aircraft, airplane, person* or *thing*, bounded regions like *place, area* and *aerodrome areas* also fall within a large group of entities, but are labeled as locations according to the terminology of FrameNet. Activities like *flying, take-off, movement* and *transport* are typically present in the field of aviation, as well as a number of procedures that we ontologically define as a type of an activity. Examples (1) to (5) show the definitions of five aviation concepts defined within the semantic frames Aerodrome, Air_transport and Flight, in which frame elements are marked in Italics. Figure 1. shows how sentences were annotated in WebAnno.

(1) aerodrome – a defined *area* on land or water, including any *objects, installations and equipment*, intended to be used either wholly or in part for the movement, take-off, landing and parking of *aircraft*.

(2) *airline* – an *air operator* that uses *aircraft* to *transport* people and/or goods for *commercial purposes*.

(3) *air transport* – the *transport* of *people* or *things* from one place to another by means of an *aircraft*.

(4) *flight* – the *flying* of *aircraft* from any *aerodrome* to a *destination* aerodrome.

(5) *take-off* – an *aircraft operation* during which an *aircraft* accelerates from the stop phase, leaves the ground and reaches the required the *flight level*.

Annotation	
3	zrakoplov – letjelica teža od zraka s vlastitim pogonom koja uzgon u letu dobiva poglavito aerodinamičkim reakcijama na površinama krila koja u svim uvjetima leta ostaju nepokretna.
4	let – odlazak s određenoga aerodroma prema određenome određišnom aerodromu.
5	zračni promet – letenje i kretanje zrakoplova po aerodromskim površinama za kretanje.
6	kontrola zračnog prometa – usluga uspostavljena radi sprečavanja sudara između zrakoplova ili između zrakoplova i prepreka na manevarskoj površini te radi provedbe i održavanja redovitoga protoka zračnoga prometa.
7	uzlijetanje – zrakoplovna operacija tijekom koje zrakoplov ubrzava iz stanja mirovanja na početku zaleta, odvaja se od tla i postiže propisanu visinu.
8	slijetanje – zrakoplovna operacija tijekom koje se zrakoplov spušta s određene visine, dodiruje uzletno-sletnu stazu, smanjuje brzinu i potpuno se zaustavlja.
9	zračni prijevoz – prijevoz osoba ili stvari zrakoplovom iz jednoga mjesta u drugo .
10	zračni prijevoznik – zračni operator koji komercijalno zrakoplovima prevozi osobe i/ili stvari.

Figure 1. Annotated frame elements as represented in Croatian definitions extracted from Struna

The parallel English-Croatian aviation corpus compiled within the project was used to extract sentences used for annotation. The corpus is compiled from the Directory of legal acts of the European Union, from the chapter "Transport policy", subchapter Air transport in English and Croatian. Out of 220 documents from the "Air transport" subchapter, 178 legal acts are taken having both language versions. The texts are downloaded from the EUR-Lex database, and entered into the Sketch Engine's corpus compilation module. The corpus was queried for the target terms *aerodrome*, *aircraft*, *airline*, *airspace*, *air traffic*, *air traffic control*, *air transport*, *flight*, *landing* and *take-off*, and sentences were manually extracted and annotated. Examples (6) to (9) show some of the sentences with the target terms *aerodrome*, *aircraft*, *airline* and *flight*. Semantic roles are marked in small caps in square brackets following the sentence element they label.

(6) The flight previously notified by a basic flight data process [THEME] will now not enter the airspace of the notified unit [FINAL_LOC].

(7) An applicant [AGENT] shall fly the aircraft [THEME] from a position where the PIC functions can be performed [INITIAL_LOC] and to carry out the test [THEME] as if there is no other crew member [SETTING].

(8) An aircraft taxiing on the manoeuvring area of an aerodrome [AGENT] shall give way to aircraft taking off or about to take off [BENEFICIARY].

(9) Similarly, the airline operating the aircraft [PIVOT] needs underlying economic authority [THEME] from the DOT [SOURCE].

The terminology of the semantic roles in the LIRICS tagset obviously differs from FrameNet's roles in some aspects, but differences could be noticed between certain roles in other projects, e.g. VerbNet and PropBank. E.g., Initial location and Final location as used in LIRICS and in our annotation correspond to Source and Destination in VerbNet, while the Setting corresponds to the meaning labeled as Circumstances in FrameNet.

4. Discussion

Since semantic relations are crucial in connecting frame elements, and consecutively in determining the relations between different frames, they are reflected at the sentence level as verbs, their arguments and adjuncts, to which semantic roles are assigned. The number of semantic roles and their definitions vary depending on the degree of abstractedness and concreteness. Roughly, there are two approaches to semantic role labeling. The first approach does not have roles specific to one class, e.g. LIRICS (Petukhova & Bunt, 2008) and VerbNet (Kipper et al., 2008) with hierarchically organized semantic roles. The second approach, developed within FrameNet has more than 2000 frame elements or semantic roles that are verb-specific and frame-specific. Gantar et al. (2018), following and simplifying the Prague Dependency Treebank (PDT), use 25 semantic roles for the annotation of examples extracted from the general language corpora in Slovenian and Croatian. An even more reduced tagset (17 semantic roles) is applied in the creation of the semantic layer of the Croatian Dependency Treebank (Farkaš, Filko, & Tadić, 2016). Semantic role labeling in specialized knowledge resources differs to some extent, depending largely on the nature of conceptual relations within a particular domain. E.g., the semantic roles set used in Ecolexicon (Araúz et al., 2012) is based on the organization of the field of environment around *event* as the key category of the domain as well as the roles applied in the framed version of the Canadian database DiCoEnviro (L'Homme, Robichaud, & Rüggeberg 2014).

An ideal terminological intensional definition should contain a concept superordinate to the one that is being defined, and the delimiting characteristics, i.e. the characteristics that set out the defined concept from concepts similar or related to it. Ontological categories labeled in the analyzed definitions refer to frame elements that invoke particular semantic frames that structure the field of aviation. If we take the frame of Air_Transport as an example, it is defined as 'the transport of people or things from one place to another by means of an aircraft'. Transport, People, Things and Aircraft can be said to be the frame elements invoking this frame, and each one of them is then defined and put into relation with other semantic

frames they might also be part of. However, the relations among different frame elements in a frame are only implicitly present in the frame definition. A better insight into the complexity of both intraframe and interframe relations is gained at the lexical level, where frame elements are expressed as semantic roles.

Choosing a model and a specific tagset to be applied in semantic role labeling depends largely on the theoretical approach referred to. Although FrameNet offers an exhaustive list of roles, its magnitude also presents its largest drawback, i.e. a difficulty to correlate with resources that apply a more general and limited SRL tagset. Using less semantic roles in annotation, on the other hand, can lead to leaving out useful semantic information. Duration is thus a semantic role missing in the LIRICS tagset, but it is an important role for labeling examples from an aviation corpus because it marks a relevant component in activities like aircraft procedures. However, a more thorough analysis still needs to be carried out in order to reach a conclusion on the benefits and pitfalls of adding more specific semantic roles, as well as on the methodology applied.

5. Conclusion

Online terminological resources that refer to the organization and presentation of specialized knowledge according to dynamic categories like semantic frames mostly refer to FrameNet's methodology in defining semantic frames and frame elements. Some of them, like the Ecolexicon database, set up on the principles of Frame-based terminology (Faber et al. 2012), and the Canadian databases DiCoEnviro and DiCoInfo (L'Homme, Robichaud, & Rüggeberg 2014) have modified and simplified the FrameNet's semantic roles. Although the merging of resources and the interchange of data is a pressing need in contemporary digital lexicography, the purpose of each resource and the needs of its users must bear more relevance in deciding what methodology to apply. The LIRICS semantic tagset that was applied in the annotation of sentences from aviation related texts has to be enriched with a few semantic roles that are not present in the current list, e.g. Duration and Inanimate agent. However, the tagset should not be too large, otherwise it would slow down the process of annotation, without contributing much to the project goals. Categories of specialized knowledge can be defined in terms of semantic relations that bind them together, but it must nevertheless be done in a precise, clear and concise manner so to reflect both terminological needs and the need for good terminology.

References

- Araúz, P. L. et al. 2012. Specialized language semantics. In P. Faber (Ed.), *A Cognitive Linguistics View of Terminology and Specialized Language* (pp. 95-176). Berlin/New York: De Gruyter Mouton.
- Brač, I., M. Bratanić, & A. Ostroški Anić. (2015). Hrvatsko nazivlje i nazivoslovlje od Šuleka do Strune - hrvatski jezik i terminološko planiranje. In M. Bratanić, I. Brač, & B. Pritchard (Eds.), *Od Šuleka do Schengena: terminološki, terminografski i prijevodni aspekti jezika struke* (pp. 3-26). Zagreb/Rijeka: Institut za hrvatski jezik i jezikoslovlje & Pomorski fakultet u Rijeci.
- Bratanić, M., & A. Ostroški Anić. (2013). The Croatian National Termbank STRUNA: A New Platform for Terminological Work. *Collegium Antropologicum*, 37(3), 677-683.
- Eckart de Castilho, R. et al. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka, Japan* (pp. 76-84). Retrieved from <https://pdfs.semanticscholar.org/aee7/1bfa28ec1bad78d4bd4aadcab168aa6b3b13.pdf>
- Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.
- Farkaš, D., M. Filko, & M. Tadić. (2016). HR4EU – Using Language Resources in Computer Aided Language Learning. *CLIB 2016 Proceedings* (pp. 38-46). Retrieved from https://bib.irb.hr/datoteka/852491.clib_farkas.pdf
- Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm* (pp. 111-137). Seoul, South Korea: Hanshin Publishing Company.
- Fillmore, C. J., Christopher R. J., & M. R. L. Petruck. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), 235-250.
- Gantar, P. et al. (2018). Towards Semantic Role Labeling in Slovene and Croatian. *Konferenca Jezikovne tehnologije in digitalna humanistika, Ljubljana*. Retrieved from http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Gantar-et-al_Towards-Semantic-Role-Labeling-in-Slovene-and-Croatian.pdf
- ISO 704. (2000). *Terminology work – Principles and methods*. International Organization for Standardization.
- Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.
- Kipper, K. et al. (2008). A large-scale classification of English verbs. *Lang Resources & Evaluation*, 42, 21-40.

L'Homme, M. C. (2012). Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, 23, 233–252.

L'Homme, M. C, B. Robichaud, & C. S. Rüggeberg. (2014). Discovering Frames in Specialized Domains. In N. Calzolari et. Al (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1364-1371). Reykjavik, Iceland: European Language Resources Association (ELRA).

Melby, A. K. (2015). TBX: A terminology exchange format for the translation and localization industry. In H. J. Kockaert & F. Steurs (Eds). *Handbook of Terminology, Volume 1* (p. 392-423). Amsterdam/Philadelphia: John Benjamins.

Petukhova, V., & H. Bunt. (2008). LIRICS semantic role annotation: design and evaluation of a set of data categories. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 39-45). Retrieved from <https://pdfs.semanticscholar.org/732c/65885e1e664c44db6ad723425547633dad7a.pdf>

Ruppenhofer, J. et al. (2010). FrameNet II: Extended Theory and Practice. Retrieved from <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>