Corpus-based Lexicography for Sesotho

by

Mmasibidi Setaka

Student Number 10657542

A dissertation submitted in fulfilment of the requirements for the degree Master of Arts

in the Department of African Languages at the
Faculty of Humanities UNIVERSITY OF PRETORIA,

SUPERVISOR: Prof. D.J. Prinsloo

August 2018

Acknowledgements

Thanks God for allowing this opportunity and forever gracing me with your mercy and blessings. To my queen mama Ellen Setaka and my sister Madibe Setaka and brother Ntsheke Setaka, you have been more than a blessing in my life and I am most grateful for your unwavering support, advices and encouragement. You make my world brighter. To my son Relebohile Setaka, you motivate me to do more. To my late grandparents, I know you are watching over me always. Thank you and Rest in Peace.

Prof. Prinsloo, thank you for believing in me. Your contribution in my studies gave me life lessons about patience, hope and success. I cannot thank you enough for all that you have done for me.

SADiLaR, all could have been really hard to achieve if it was not for you. Thank you for the opportunity. Pieter Labuschagne, thank you for always helping me.

To my prayer warrior, Matlhogonolo Tsae, thank you for your prayers and constantly enquiring about my studies. That is true friendship and I am most grateful to have you in my life.

To all my friends and family who stood by me, I salute you.

Modimo o le hlohonolofatse.

Table of Contents

Abstract

For centuries, dictionaries were compiled based upon the knowledge of the lexicographer and information retrieved from manually consulted sources, mainly through a process of reading and marking. This approach meant that much of the information used in the dictionary relied upon the knowledge of the lexicographer. It is vital to rely on the lexicographer’s knowledge of the language but this has its shortcomings, since there is no single individual who knows all the words or terms, their meanings and usage, the words they combine with, and so on, in a specific language. The utilization of this method left room for errors and omissions because the lexicographer could easily overlook some words due to factors like time, fatigue, limited knowledge of the lexicographer, etc. Important words, for example words likely to be looked for by the target users of the dictionary, could accidentally be omitted. In the 1980s, the corpus era was born and the lexicography field changed forever. Collins COBUILD in Birmingham spearheaded this era with the publication of the first corpus-based dictionary, the Collins COBUILD Dictionary in 1987. Since the corpus era began, lexicographers no longer rely solely on their knowledge of the language, intuition, or the limited information gathered from available written sources, which are very limited for African languages. The corpus allows the lexicographer to have access to huge volumes of authentic data from written texts and transcribed oral data. This research will therefore critically discuss dictionary compilation for Sesotho and spearhead the use of corpora in the compilation of Sesotho dictionaries, so that lexicographers do not compile dictionaries as if they are compiling the first dictionary for the language. In addition, they should take into account tasks like lexicographic planning, amongst other factors required to compile a good user-friendly dictionary.

Key words

Corpora, collocations, concordances, lexicography, lexicographical planning, microstructure, macrostructure, lemmatisation.

Chapter 1: Introduction

1.1 Introduction

For centuries, the process of compiling a dictionary was based largely on the knowledge of the lexicographer and information retrieved from manually consulted sources. Lexicographers supported their intuition and knowledge of the language through reading and marking. Much of the information used in the dictionary thus relied on the knowledge of the lexicographer. Although it is vital to rely on the lexicographer's knowledge of the language, it is not possible for any single individual to know and master all the words and terms, their meanings, uses and nuances, even if that individual is an unparalleled expert of lexicography. There are possibilities that a lexicographer can overlook or miss some words or terms. This means that important words - defined in this dissertation as words most likely to be looked for by the target user - will most likely be omitted and therefore not be lemmatised in the dictionary. The process of reading and marking, which essentially involves obtaining information from manually consulted sources, also meant that the lexicographer can overlook some words due to constraints such as time, fatigue, insufficient space in the dictionary, etc., because this process is, by its nature, laborious and time consuming. The lexicographer, for instance, had to manually consult and check multiple sources before reaching a conclusion. In addition, the selection of lemmas (headwords) for a dictionary was conducted by lexicographers using only their intuition and they would often enter words into the dictionary as they encountered them. Snyman et al., (1990: preface) explain this notion as follows:

> The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compiling of a dictionary. This can take place simply because the lexicographer had not encountered such words. We can only hope that there are not too many examples of this kind.

Relying on the intuition of the lexicographer for lemmatisation of words is no longer seen as an efficient (or even wise) strategy in the corpus era. In fairness, the intuition of the lexicographer is informed mainly by their knowledge of the language and other subjective factors, and this surely cannot be fair for a process as crucial as the identification, analysis and treatment of lemmas. Instead, it will expose the lemmatisation process to vulnerabilities such as human error, omission of some words, etc. In other words, the lexicographer will not lemmatise important words that were most likely to be looked for by users, just because he/she has not come across them and/or does not know them. In this way, the lexicographer would not be fulfilling their job, which is to lemmatise words that are most likely to be used, especially for African languages that have limited recorded sources. This would mean that, if a lexicographer had not encountered a word, it would not be lemmatised. This also means that even the most common and basic words could be accidentally omitted from the dictionary. This will, inter alia, compromise the quality of the dictionary, and it will be riddled with other weaknesses such as omission of certain important words and an incomplete list of key categories/classes of words, all because of the lexicographer's lack of information.

Another important aspect to consider is the users, and Gouws and Prinsloo (1998:18) emphasize this point, thus:

> The user-perspective, so prevalent in modern-day metalexicography, compels lexicographers to compile their dictionaries according to the needs and research skills of well-defined target user groups. The dominant role of the user has had a definite effect on the compilation of dictionaries as well as on the evaluation of their quality. Good dictionaries do not only display a linguistically sound treatment of a specific selection of lexical items. They are also products that can be used as linguistic instruments by their respective target user groups. The better they can be used, the better dictionaries they are.

This means that a lexicographer has to have a certain group of users in mind when compiling a dictionary because without them the publication will lose focus and context. For example, if a lexicographer decides to compile a medical dictionary, it will be a dictionary for special purposes meant for a certain group of specialists and it will not be helpful to a high school student who is trying to find information for their Sesotho grammar school project. In the 1980s, however, the corpus era dawned upon lexicography, spearheaded by Collins COBUILD in Birmingham with the publication of the first corpus-based dictionary, the COLLINS COBUILD DICTIONARY in 1987 (Collins, n.d.). This ground-breaking work revolutionized and changed the principles and practice of lexicography, most likely 'forever', and also meant that a lot of problems previously encountered by lexicographers were solved - including the issue of information retrieval. A corpus can be defined as a collection of texts and speech of a particular language and represents actual language usage amongst people in a given setting.

Laviosa (2002:34) defines a corpus as follows:

> A corpus is generally referred to as either a collection of texts or a collection of pieces of language. Both definitions express an important feature of a corpus, namely that it is a sample of texts, either full running texts or text extracts assembled according to explicit design criteria.

Since the corpus era began, lexicographers no longer have to rely solely on their own knowledge of the language or intuition and the limited information that is gathered from available written sources, which are very limited for African languages. The compilation and availability of a corpus means that lexicographers now have access to huge volumes of authentic data from written texts and transcribed oral data. Studying the types and frequency of words in the languages, as well as the contexts in which they occur, opened new opportunities for lexicographers:

> For ages lexicographers battled to increase the quality of dictionaries. The corpus, however, has suddenly opened up new horizons for dictionary makers. (Prinsloo and De Schryver 2000:188)

We have to applaud the compilers of the first corpus dictionary because they uplifted lexicographic standards in many ways, and African languages need to step up and take advantage of this new and very helpful tool, which has already revolutionized many languages and improved the process of the collection of data, processing and compilation of dictionaries. The advent of the corpus era means that lexicographers will no longer consume much of their time trying to manually gather words on introspection and to rely solely on their own knowledge for meaning description, usage, sense distinctions, etc. Corpus data substantially enhance the quality of dictionaries. The time has come for scholars of Sesotho and other African languages to produce well-structured and lexicographically-planned dictionaries with all frequently-used lemmas.

Utilization of corpus data impacts on the two main areas of dictionary compilation, viz: macrostructural and microstructural activities. These constitute the core of research in this dissertation. On the level of the macrostructure, lexicographers no longer rely on their intuition in respect of the selection of lemmas, their task instead is to utilize sound lexicographic strategies for the selection of lemmas. Microstructure, on the other hand, distinguishes different senses of a word, constructs examples of usage, enhances the quality of treatment of lemmas and improves quality of definitions, among other things.

1.2 Rationale and background to the study

Dictionary compilation for African languages, and Sesotho in particular, does not stand or develop in isolation, but is influenced by trends and changes in international lexicography. In the past four decades, three main phases of development are distinguishable, and these are as follows: the Zgusta era, the Wiegand era, and the Function Theory of Bergenholtz and Tarp. These are described in more detail by Gouws and Prinsloo (2005) and others. First to emerge was the so-called Zgusta era in the 1970s where the focus was on dictionary content, and dictionary compilation was regarded as a linguistic activity. Then came the 1990s trend, often described as the Wiegand era, where the emphasis moved to the structure and structural components of dictionaries. Finally, in the current era, the Function Theory of Bergenholtz and Tarp takes centre stage. The Function Theory's main focus is on the user-perspective, particularly in respect of the compilation of user-friendly dictionaries. In other words, the production of dictionaries that satisfy the cognitive, text/speech reception, text/speech production and communicative needs of a set of narrowly defined target users.

Gouws and Prinsloo (2005:3-8) summarise these phases of development as follows:

> Soon after the publication of the *Manual of Lexicography* the influence of Zgusta's ideas was already noticeable, resulting in the rapid growth of theoretical lexicography but also in an improvement in the quality of new dictionaries… The Wiegand era has been characterised by the identification of the different components of dictionary articles and by a meticulous description of their specific structure and function… The profile of the users must be determined, and eventually the relation between the needs of each type of user in each type of user

> situation and the data included in a dictionary to satisfy these needs constitute the basis for the theory of lexicographic function.

These three stages of development coincide with three stages of practical dictionary compilation which are influenced by technological development, in particular the dawning of the computer age and the birth of fields like Human Language Technologies (HLT) and Natural Language Processing (NLP). The utilization of corpora is in principle nothing new to lexicography – the compilers of the Oxford English Dictionary (OED) have been using teams of assistants for many years to do reading and marking of English literary works to be used in the selection and presentation of lemmata (macrostructure), distinction of senses, construction of examples, and identifying typical collocations (microstructure). The computer age enabled the storage of corpora on convenient electronic devices like computers, which in turn enabled computer query software programs such as *WordSmith Tools* to manipulate the corpus data within a matter of seconds as opposed to time-consuming reading and marking activities.

Manipulation of corpus data, firstly, render alphabetical and frequency word lists as well as representation of keywords in context for utilization on the macrostructural and microstructural level for dictionary compilation. Secondly, such outputs can be used for more advanced applications and dictionary compilation tools such as lexicographic rulers and block systems that regulate balance in the dictionary in terms of the appropriate size of each alphabetical stretch in the dictionary. Thirdly, corpus query outputs are utilized for sophisticated applications such as word sketches where the entire behavioural pattern of a word is summarized, often on a single page. These issues will be discussed in the relevant chapters.

The computer era has thus enabled a new approach to dictionary compilation: corpus-based dictionaries. This era also opened the door to dictionaries on CD-ROM (Computer Disc Read-Only Memory) and such dictionaries were developed in parallel to print or paper dictionaries. As a result, "developments in the field of lexicography saw numerous dictionaries being produced on CD-ROM and on the internet," (Gouws and Prinsloo 2005:8). The past two decades also experienced the growth of the internet and introduced a new medium in dictionary compilation, namely electronic/online dictionaries (EDs). The electronic era was met with great enthusiasm and expectations. Early publications on EDs were all about the potential of the new medium and the expected revolution it would bring along, such as antiquating the paper-based dictionary in a decade or two. Hence, Prinsloo (2001) and De Schryver (2003) believe that the electronic dictionary (whether on CD-ROM, online, or handheld), would supersede the paper based dictionary in ways unimaginable in the paper-dictionary dimension, just as the computer had completely superseded the typewriter. With this idea in mind, De Schryver titled his ground-breaking work as '*Lexicographers' Dreams in the Electronic-Dictionary Age'*.

Electronic dictionaries developed much slower than expected, relatively speaking, but regained momentum in the past few years and are regarded currently as the new attraction at major international conferences. Hence Gouws and Prinsloo (2005:8) pointing out that:

> The last decade has witnessed tremendous developments in the field of electronic dictionaries. The electronic medium has become increasingly important for the transfer of knowledge and lexicography had to respond to this.

The holy grail in electronic dictionary development currently is the development of interactive dictionaries that link electronic dictionaries, different types of processed and unprocessed corpora, and internet data. The interested reader is referred to Wanner et al. (2013), Alonso Ramos et al. (2014), Verlinde et al. (2010), Prinsloo et al (2012, 2014 and 2017). Granger and Paquot (2012:2) summarise the developments brought by this era as follows:

> Things started to accelerate … with the rapid development of a range of new mediums, in particular handheld devices and (a little later) online dictionaries.

African language lexicography, and especially Sesotho lexicography, was slow to react to the opportunities that were brought by the corpus era, particularly with regard to the development of electronic dictionaries. Prinsloo is acknowledged for starting the corpus era in Africa with a pioneering publication "*Towards computer-assisted word frequency studies in Northern Sotho*", published in 1991, and for his actual compilation of the first corpora for African languages. The eleven (11) National Lexicography Units (NLUs) in South Africa use corpora for dictionary compilation for all the official languages of South Africa. However, in the case of Sesotho the compilation of dictionaries continues, to a large extent, without the utilization of corpora and thus reflects lexicographic practices of the pre-corpus era. The view expressed by De Schryver and Prinsloo (1999:1) is particularly relevant for Sesotho:

> … today's compilers fail to seriously take into consideration the importance of a corpus as a useful tool for the description of actual language usage. They have to understand and know that corpora have a lot to contribute towards the compilation of modern dictionaries, thus they cannot continue to ignore them.

The situation for Sesotho lexicography in general is much worse than even that of its sister languages, Sepedi and Setswana, in the Sotho group of languages. According to Mojela (2007), Prinsloo (2009), and De Schryver and Prinsloo (2001) lexicographic research for Sesotho is almost non-existent compared to Sepedi for which numerous aspects, mainly in respect of lemmatisation, have been researched and implemented in dictionaries. Likewise, for Setswana much work has been done by scholars such as Sebolela (2009) and Otlogetswe (2013a and 2013b). The main objective of this dissertation is to lay the foundation and to show the way for the compilation of corpus-based dictionaries of high lexicographic quality for Sesotho.

The compilation and utilization of electronic corpora for Sesotho lexicography will enable the compilers of paper-based as well as electronic Sesotho dictionaries to bring Sesotho lexicography on par with the best of modern dictionaries and to facilitate the enhancement of its lexicographic standards.

Although there have been commendable efforts in African-languages lexicography and dictionary compilation in recent years, there are still gaps. In other words, the quality in terms of standard lexicographic principles remains inadequate and a headache for both users and scholars. Having studied dictionaries for several African languages, Gouws (1990:55) summarizes their status as follows:

> The majority of dictionaries for African languages are the products of limited efforts not reflecting a high standard of lexicographical achievement ... with a few exceptions these dictionaries offer only restricted translating equivalents and reflect a complete lack of lexicographical planning.

In African languages, a complex interplay exists between lexicographic traditions and lemmatisation strategies. These are further complicated by tension between conjunctive and disjunctive orthographies of various African languages and problematic aspects of lemma identification influenced by these spelling systems or conventions, i.e. conjunctive spelling system versus disjunctive spelling system. For example, a lexicographer has to deal with different lemmatisation traditions, approaches and a number of lemmatisation problems. Prinsloo (2009) provides an extensive discussion about this phenomenon. The current African language lexicography is also heavily influenced by certain lexicographic traditions and global trends and changes – such as the user perspective and the emphasis on user-friendliness, a more functional approach, increased attention to dictionaries suitable for text production, corpus-based dictionaries superseding the traditional introspection-based dictionaries, etc.

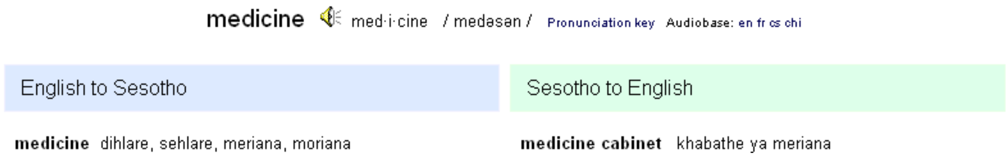
It could also be argued that the electronic era dawned upon African language lexicography at a time when the goal of compiling good dictionaries had not yet been attained. In many cases answering the call for electronic dictionaries resulted in word lists with one or more translation equivalents being put on the internet. Prinsloo (2012:129) explains:

> Not much progress has been made in terms of the utilization of the virtues of the electronic medium such as the great capacity and speed characteristic of electronic products, combined with enhanced query and data retrieval technology. Furthermore, not much is exhibited in terms of innovative features characteristic of online dictionaries for English, e.g. pop-up access, bringing together of related items, new routes to the data, less dependency on alphabetical order, fuzzy spelling, intelligent extrapolation of characters keyed in, audible pronunciation, etc.

Generally speaking, African languages are poorly resourced. Characteristics of electronic dictionaries (EDs) for lesser-resourced languages include the following: limited size of the dictionary, small number of lemmas treated, limited number of data types, which is often nothing more than word lists with one or two translation equivalents, etc. The electronic dictionary situation in Africa is emblematic of the state of lesser-resourced languages of the world. Case studies of these languages, and especially the lesser-resourced languages spoken in South Africa, reflect the typical situation of African-languages paper dictionaries: their entries simply consist of L1 head word lists with L2 translations. In other words, the nine official African languages reflect the current situation in lesser-resourced languages. Vetulani, Uszkoreit, Kubis (2016:401) capture this worrying situation as follows:

> There are more than 6,000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies.

Electronic dictionaries for Sesotho are limited to mere word lists or to word lists with one or more translation equivalents added on. Consider the following example:

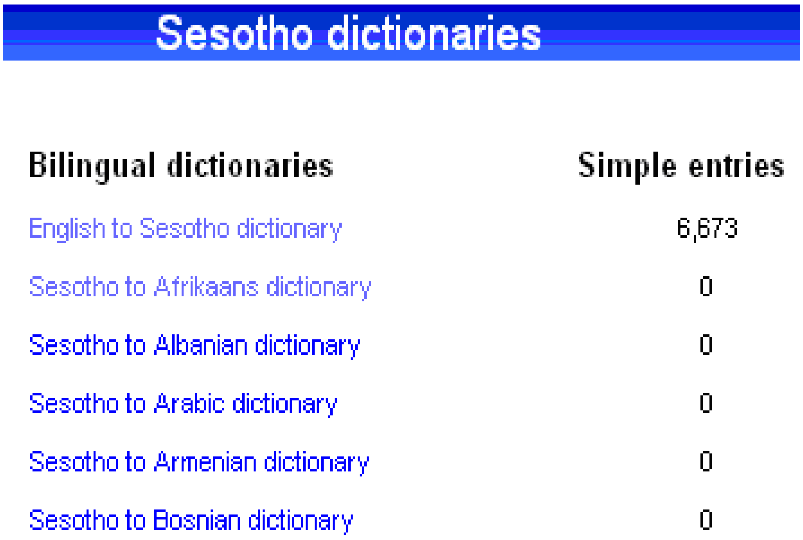


medicine med·i·cine / medəsən / Pronunciation key Audiobase: en fr cs chi

English to Sesotho	Sesotho to English
medicine dihlare, sehlare, meriana, moriana	**medicine cabinet** khabathe ya meriana

Figure 1.1: The treatment of medicine in Dicts.info

Compare the following example where Sesotho is simply drawn into existing online dictionary networks and mathematical combinations of possible dictionaries, where the aim seems to be to collect as many languages as possible instead of compiling quality electronic dictionaries. *Webster's Online Dictionary* calls itself 'Earth's largest dictionary with more than 1,226 modern languages and Eve!'. "*Dicts.info* offers an English-to-Sesotho dictionary and lists no less than 55 empty bilingual dictionaries with Sesotho as source language" (Prinsloo 2013:252). Consider the following:



Sesotho dictionaries

Bilingual dictionaries	Simple entries
English to Sesotho dictionary	6,673
Sesotho to Afrikaans dictionary	0
Sesotho to Albanian dictionary	0
Sesotho to Arabic dictionary	0
Sesotho to Armenian dictionary	0
Sesotho to Bosnian dictionary	0

Figure 1.2: Bilingual Sesotho dictionaries (http://www.dicts.info/dictlist1.php?l=sesotho)

> These dictionaries are technically fully functional empty shells ready to be populated with lexicographic content. One can even click on the Sesotho to Afrikaans dictionary and enter the Sesotho word *hore* ‘so that’ in the search box, a search is performed and not surprisingly returning a ‘not found’ result. Prinsloo (2012:130)

This dissertation will illustrate that a substantial amount of work needs to be done to eventually have sophisticated electronic dictionaries for lesser-resourced languages.

For both paper and electronic Sesotho dictionaries, the conclusion of Gouws (1990:53) is still relevant: that “currently available dictionaries are the products of limited efforts not reflecting a high standard of lexicographical achievement.” Typical examples in support of the above-mentioned statements are inconsistencies found on the macrostructural level and inferior treatment of lemmas on the microstructural level of these dictionaries. Many of these inconsistencies can be attributed to the failure to use electronic corpora, written as well as spoken, that became available to lexicographers in the past two decades. If African languages lexicography is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should become an absolute priority, cf. De Schryver and Prinsloo (2000a and 2000b) and Prinsloo and De Schryver (1999 and 2001b). Prinsloo and De Schryver (2000) also note that African-languages dictionary compilers have an even greater challenge since they are mediators between a complicated grammatical system on one hand and the often-inexperienced dictionary user on the other.

1.3 Aims of the study

The aims of the study can be divided into four categories as follows:

1. to give a perspective on Sesotho in terms of its position as an African language and within the Bantu-language family in particular,
2. to give a broad survey of core aspects of dictionary compilation,

3. to give a critical evaluation of African-language dictionaries with specific focus on Sesotho dictionaries, and
4. to provide recommendations and attempt to lay the foundation for the compilation of corpus-based dictionaries of high lexicographic quality for Sesotho focusing on macrostructural and microstructural aspects.

Aspects of the aims outlined above will now be presented in more detail.

1.4 Building organic corpora for Sesotho

Electronic corpora can simply be defined as large collections of text (it can be in Sesotho or any other language) stored in electronic format. A corpus is then used as input into so-called corpus query programs that manipulate the data into formats such as word lists (alphabetical and frequency) and concordance lines (giving a word in contexts, with co-text to the left and to the right of the word).

One of the most interesting approaches to corpus compilation is that of the British lexicographer, B.T. Sue Atkins. At the height of the debate on balanced versus representative corpora, she introduced the concept of organic corpora:

> A corpus builder should first attempt to create a representative corpus. Then this should be used and analysed, and its strength and weaknesses identified and reported. In the light of experience and feedback the corpus is enhanced by the addition or deletion of material and the circle is repeated continually. This is the way to approach a balanced corpus. One should not try to make a comprehensive and watertight listing […] rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language […] in our ten years' experience of analyzing corpus material for lexicographic purposes, we have found any corpus – however unbalanced – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. (Atkins 1997, oral communication at Salex'97).

This approach fits the Sesotho situation perfectly because it renders roughly the same results in terms of balance and representativeness for languages where sources are much less available than for English and other well-developed languages. Prinsloo (2015) conducted an in-depth research on maximal lexicographic utilization of limited corpora and came to the conclusion that even a very limited corpus consisting of one million words (tokens) can be a very useful lexicographic tool. Frequency lists culled from these corpora will be used for evaluation and the compilation of lemma lists on the macrostructural level as well as for frequency indication in the dictionaries themselves – a modern trend in lexicography in terms of aims 1. and 2. above. Concordance lines will form the backbone of the evaluation and compilations on microstructural level in terms of the aims formulated above.

1.5 Objectives of the research

The objectives of this research are outlined and include the following: description of principles and practice for the compilation of modern Sesotho dictionaries and evaluating selected English dictionaries; critical evaluation of African-language dictionaries with specific focus on Sesotho dictionaries; and, designing a comprehensive lexicographic approach for the compilation of Sesotho dictionaries.

1.5.1 Description of principles and practice for the compilation of modern Sesotho dictionaries and evaluating selected English dictionaries

The first objective is to describe the principles and practices that are characteristic of the compilation of modern dictionaries and then to evaluate according to these principles a selection of major English dictionaries, the so-called “Big Five”. These are Cambridge Advanced Learner's Dictionary (CALD), Collins COBUILD Advanced Dictionary of English (COBUILD), Longman Dictionary of Contemporary English (LDOCE), Macmillan English Dictionary for Advanced Learners (MED) and Oxford Advanced Learner's Dictionary of Current English (OALD). The focus will be on material collection and corpus building, frame structure, access structure, macrostructural aspects and microstructural aspects. Specific emphasis will be on macrostructural aspects such as the compilation of the lemma list in the central text, utilization of frequency counts, selection strategies, and so on, because they are regarded as the most common shortcomings in African Languages. On the level of the microstructure, the focus will be on sense distinction, selection of examples and, most importantly, definitions (paraphrase of meaning) of a high lexicographic standard.

1.5.2 Critical evaluation of African-language dictionaries with specific focus on Sesotho dictionaries.

The principles and practice that are characteristic of the compilation of modern dictionaries and the knowledge gained from the evaluation of dictionaries of major languages of the world will be applied in a critical evaluation of African-language dictionaries, with specific focus on Sesotho dictionaries. This evaluation will focus on the same macrostructural and microstructural aspects as given in 1.5.1 above, i.e., compilation of the lemma list, utilization of frequency lists etc. on the macrostructural level, and on microstructural level the focus is on sense distinctions, selection of examples of usage, etc., with special attention to lexicographic aspects unique to African languages such as tonal indication and lemmatisation strategies. Critical analysis of Sesotho dictionaries available currently will be done with special reference to Mabille and Dieterlen (1950), Batho Hlalele (2005) and Bukantswe Online dictionary.

A detailed evaluation of existing Sesotho dictionaries will be performed in terms of the following macrostructural and microstructural aspects:

- Coverage of frequently-used words
- Effective use of dictionary space
- Use of lexicographic examples

1.5.3 Designing a comprehensive lexicographic approach for the compilation of Sesotho dictionaries

The objective here is to utilize the research outcomes in 1 and 2 with the focus on the compilation of macrostructures and microstructures of high lexicographic standards for both monolingual and bilingual dictionaries of Sesotho.

Conclusions reached thus far from the study of core literature, by amongst others Zgusta (1971), Landau (2001), Gouws and Prinsloo (2005), and preliminary analysis of Sesotho dictionaries indicate that Sesotho dictionaries have the following shortcomings:

- the lack of proper lexicographic planning;
- the absence of written and spoken corpora for Sesotho and corpus querying for macro- and microstructural application;
- over-emphasis on the comment on form, e.g., grammatical information;
- insufficient semantic information;
- imbalances in alphabetic categories, i.e., over- versus under-treatment;
- deviation from a normal alphabetic ordering;
- inconsistency in lemmatisation, e.g., stem and word lemmatisation within the same dictionary;
- lack of a selection strategy for lemmas; and
- absence of user feedback.

The envisaged required lexicographic activities for Sesotho are therefore extended to the compilation of corpus-based Sesotho dictionaries that can be broken down into the following phases:

- Enlarge the current Pretoria Sesotho text corpus.

- Formulation of a comprehensive compilation and revision strategy for future practical and user-friendly monolingual and bilingual Sesotho dictionaries.

1.6 Research methodology

This study is purely qualitative and literature-based.

- The study will firstly present a theoretical conspectus of the creation of electronic corpora followed by a practical exploration for African languages.

- Critical evaluation of both paper and electronic dictionaries constitutes a major part of this research and it will be done in two phases.

    - Phase 1: Evaluation of two of the so-called ‘Big Five’ English dictionaries: MED and Collins English-French Dictionary. The evaluation will be limited to the compilation of their macrostructures and treatment of lemmata in the microstructure.

    - Phase 2: Evaluation of one monolingual (Sethantšo sa Sesotho), one bilingual (Southern Sotho-English Dictionary) and one internet dictionary (Bukantswe Online) and again, evaluation will be limited to macrostructure and microstructure.

The study will also focus on various applications of lexicographic principles and practice in the broad field of lexicography.

1.7 Preview and layout of chapters

This dissertation consists of eight (8) chapters that have been arranged in the manner described below:

Chapter 1 introduces the concept of the corpus era and highlights important issues which have been solved since the dawn of the corpus era. The aim is to contrast these with the pre-corpus era dictionary compilation approaches. During that era, the process of compiling a dictionary was dependent on the knowledge of the lexicographer and information retrieved from manually consulted sources, mainly through reading and marking.

The Collins COBUILD Dictionary was the first dictionary to be compiled using corpora and its publication proved to the world that corpora have opened up new horizons and possibilities which could not have otherwise been achieved. The use of corpora has made it possible to lemmatise frequently used words and have also resulted in user-friendly dictionaries, amongst other things.

Chapter 2 will focus on the Sesotho language, an African language of the Bantu-language family in South Africa. It is part of the Niger-Congo languages which fall under the South-eastern Bantu zone. All the nine South African indigenous languages fall under this family. The nine (9) South African indigenous languages are classified into four (4) different clusters, namely, Sotho, Nguni, Tshivenda and Xitsonga clusters. For the purpose of this research, only the Sotho cluster will be analysed. This cluster consists of Sesotho, Setswana and Sepedi and, out of these three languages, Sesotho will be studied in more detail as it is the focus of this research.

Chapter 3 will focus on building or compiling corpora for Sesotho. There are three steps that take place during corpus compilation and these are corpus design, text collection and text encoding. These different stages of corpus compilation are interrelated, and one cannot exist without the other. They will therefore be discussed in great detail in this chapter.

Chapter 4 will focus on the macrostructure and how the utilization of a corpus assists lexicographers in the compilation of a lemma list, in particular with the selection of lemmas to be included in the dictionary. The corpus therefore provides sound lexicographic strategies for the selection of lemmas most likely to be looked for by the target user of a dictionary.

Chapter 5 will look at the microstructure and how the corpus has enhanced the quality of treatment of the lemma in terms of improved quality of definitions, better sense distinctions, good examples of usage, etc.

Chapter 6 will focus on the critical evaluation and analysis of the free online Macmillan Dictionary and the free Collins English dictionary. The main purpose is to identify the strategies that are used to compile these dictionaries so as to borrow them or apply) them in Sesotho lexicography and therefore help Sesotho lexicographers to follow the trends set by these English dictionaries. These dictionaries are successful in their own right and have been compiled with the use of corpora and emphasis on user needs.

Chapter 7 will critically analyse and evaluate one monolingual (Sethantšo sa Sesotho), one bilingual (Southern Sotho-English Dictionary) and one internet dictionary (Bukantswe Online). The aim is to test the veracity of the claims made by different scholars that the existing dictionaries for African languages are products of limited efforts.

Chapter 8 offers a summary of the issues relevant to corpus-based Sesotho lexicography, some conclusions, and recommendations for the way forward for Sesotho lexicography.

Chapter 2: The Sesotho language

2.1 Introduction

The African continent has between 1,500 and 2,000 languages. These languages are as diverse as their speech communities. Some are widely spoken, with speakers estimated at tens of millions, for example Swahili which has more than 100 million speakers. Other languages on the other hand have a limited spread, in terms of geography and number of speakers. For example, BBC news reported that the N|uu language has only three speakers (Fihlani, 2017).

Although the estimates for the total number of languages spoken in the continent are high, a closer analysis paints a depressing picture: most of these languages are underdeveloped, their uses being limited to elementary oral communication and other rudimentary functions.

Other countries, however, have made major strides in language development. In South Africa, for example, nine (9) of the former lesser-resourced languages have been given an official status. In other words, they have been proclaimed in the national statutes as languages that can be used legally in government communication, court proceedings, education, legislative processes, etc. Despite the complexities and other challenges that accompany a venture like this, it is a step forward for the South African indigenous languages.

Other countries in the African continent do not accord official status to their indigenous languages, while others (like in South Africa) are promoted to a national language status. This recognition is important because it promotes the use of the language(s) concerned, in both informal and formal settings.

Similarities between various African languages make some people question whether mutual-intelligible languages are fully-fledged languages or are just dialects. The classification of African languages helps us understand this linguistic phenomenon, and also distinguishes between a dialect and an independent language. This classification process groups languages into (major) families, and is used across the world – hence, for example, we have Germanic, Slavish or Celtic-language families in the case of European languages.

This chapter will begin with the classification of African languages and their spread on the continent. Then it will narrow down the scope to focus specifically on South African South-eastern Bantu Languages, with reference to the South African Constitution. Following this discussion, the focus will be on Sesotho, as it is the core of this research. With Sesotho, the focus will be on its history and classification, its dialects, sounds (i.e., vowels and consonants), its orthography, and the form and extent of its use in the news and media.

2.2 Classification of African languages

Below is the map of the African continent illustrating all the major language families and the regions where they are found. For convenience and easy reference, the regions are represented in different colours.

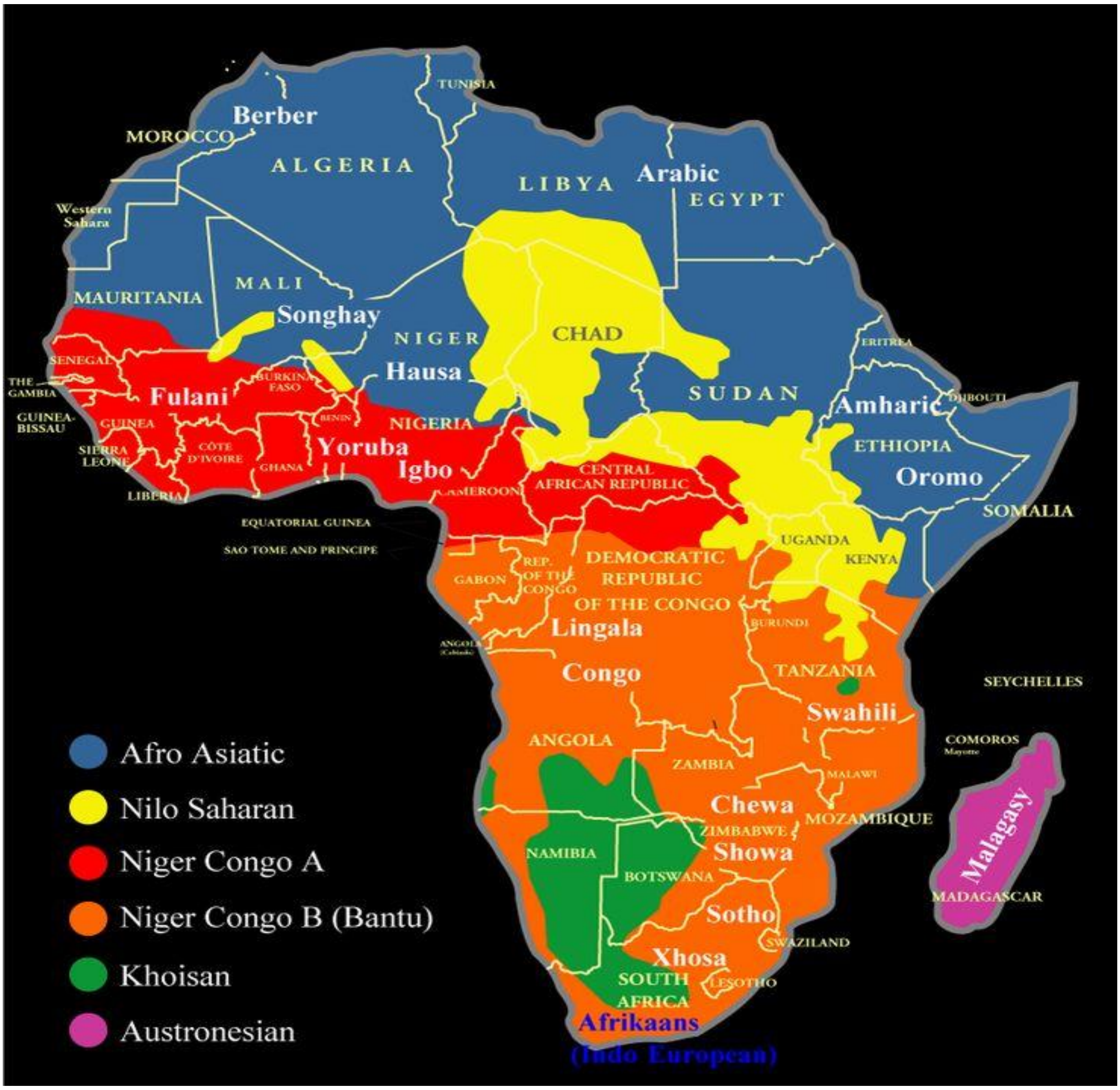


Figure 2.1: Map of Africa www.afrikanheritage.com/official-spoken-languages-of-african-countries/

African languages are divided into the following main categories: Afro-Asiatic, Niger-Congo A and B, Nilo-Saharan, Khoisan and Austronesian. According to Childs (2003: 21) the size of sub-families that constitute each major language family vary, for instance "… Niger Congo contains more than two-thirds of the 2,000 languages spoken in the continent". This major language family has between 1,350 and 1,650 languages.

Furthermore, it is not only the largest major language family in Africa, but in the world. The Niger-Congo languages are found in Western, Central, Eastern and Southern Africa and they are represented by the red and orange colours in the map above. The most widely spoken languages in Africa are Swahili (with 100 million speakers), followed by Hausa with an estimated 38 million speakers, then Igbo with 21 million, Yoruba with 20 million, Amharic 20 million and Fulani with 13 million. All these languages, together with others not mentioned, belong to the big Niger-Congo family. This assumption is supported by Collins and Burns (2007:44) point out:

> The great Niger-Congo family has six major branches and hundreds of sub-branches. The West Atlantic includes the populous Wolof and Fulbe (Fulani) originally from the coast of Senegambia. In the interior of the western Sudan are the Mande speakers – Malinke, Bambara, Sonike and the Gur languages of the peoples in the Sahel north of the West African coast, the Dogon, Mamprusi, and Mossi. Along the tropical coast of West Africa live those who speak languages of the Kwa linguistic family - Akan, Kru, Yorub and Igbo - while those speaking languages of the Adamawa-Eastern branch form a corridor of languages across Africa: from the Niger-Congo Linguistic family is Benue-Congo, which includes the Bantu group, which originated on the Cameroon-Nigeria frontier and expanded to fill the southern half of the African continent.

Furthermore, Adas (2001:235) also agrees with Collins and Burns (2007) by stating that:

> West of the Niger Bend and in the West African woodland savanna belt south of the Sudan, Niger Congo Peoples have long formed the predominant population.

The second-largest language family in Africa is the Afro-Asiatic, which has between 300 and 600 languages. This group of languages is found mainly in the northern regions of Africa, which include northern Nigeria, southern Niger, Somalia, Morocco, Algeria and Tunisia, amongst others. In the map above it is represented by the blue colour.

The other great linguistic family of Africa is Afro-Asiatic, whose speakers account for nearly a third of the languages of Africa. Today, its speakers are found in North Africa,

> the Sahara, the Nile Valley, the Middle East and Arabia and Hebrew, but also the language of the dynastic Egyptians as well as the far-flying Berber and Taureg of north Africa. (Collins and Burns 2007: 45)

Next in size is the Nilo-Saharan family with about 204 languages. The peculiar feature of this family is that most of the languages that belong to it have very small speech communities. For example, only five languages in this family have more than one million speakers. Languages spoken in this family, which are represented by the yellow colour on the map, are dispersed mainly in Eastern Africa and the North Eastern regions of Africa, which include countries like Uganda, Tanzania, Kenya, Chad, the Sudan, and the communities along the great Nile River. Languages with the biggest number of speakers, about 3.5 million to be precise, in this family are Luo and Kanuri from Kenya and Nigeria, respectively. **(**Nilo-Saharan, n.d)

Adas (2001:23) also states that "the Nilo-Saharan language extends today in a widespread remnant distribution, from the bend of the Niger in the west to the eastern side of the Middle Nile Basin in the east."

Then there is the Khoisan-language family. It is the smallest language family in Africa and is found in Southern Africa. It is highlighted in a turquoise colour on the map. This language family also has the lowest number of speakers even though it is recognized, especially in South Africa, for its rich cultural heritage. Even the motto in the South African coat of arms is written in a language that belongs to this language family. The great N|uu language also belongs to this family. Collins and Burns (2007: 45) explain this spread of the Khoisan-language family:

> The Khoisan speakers are found today predominantly in the south and south west Africa, though there are isolated Khoisan communities in East Africa and the Congo rainforest. This therefore suggests that this language family once blanketed the whole continent, even though today one of its languages includes one with just three speakers. Their numbers are insignificant compared to the rest of the peoples of Africa and include two culturally-different Southern African communities.

Finally, there is Austronesian, which is spoken mainly on the island of Madagascar and surrounding archipelagos. In the map, it is marked with purple colour.

In summary, the blue colour represents the Afro-Asiatic family of languages, which is spoken mainly in countries north of the Sahara. The Nilo-Saharan-language family, whose region is highlighted in yellow, is spoken mainly along the great Nile River in countries such as Chad, South Sudan, Uganda and Kenya in the Horn of Africa. The region highlighted in red is dominated by languages that belong to the Niger-Congo A family. This region includes countries that spread from Central African Republic to the south of Mauritania in West Africa. Most languages that are spoken in the SADC (Southern African Development Community) region belong to the Niger-Congo B family, and this family's spread is highlighted in orange. Khoisan languages are spoken mostly in Northern Cape province in South Africa, the eastern parts of Namibia, southern regions of Angola, and some parts of Tanzania. And finally Austronesian, which is highlighted in purple on the map – its varieties (languages) are spoken mainly in Madagascar and the surrounding islands.

2.3 South African Bantu languages

During the apartheid era, indigenous languages of South Africa were not given any platform or status because their speech communities were regarded as culturally inferior and therefore the languages themselves were not considered worthy to be accorded recognition and eminence. They were only used in informal settings such as homes or playgrounds. When South Africa attained democracy, it became evident that, for the country to move forward, transformation at a broader level had to take place. Therefore, the issue of language could not be ignored, especially with the rise of the black consciousness movement and the calls to include groups of previously disadvantaged people in all social and economic activities.

> The 1996 Constitution of the Republic of South Africa (Act 108) conferred official status on eleven languages; Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhosa and isiZulu. The objective of this provision is that all South Africans should be able to operate in any spoken or written official language(s) of their choice for a range of social and public contexts (Coulmas 2013:235)

This new language regime meant that formerly disadvantaged languages were also allowed to be used for legal proceedings in the judiciary, in government documents in the national administration, for legislative purposes in parliament and provincial legislatures, and also in the commercial sector, etc. It has been proposed that African languages should be introduced in education for learning and teaching purposes. Section 29(2) of the Constitution provides that everyone has the right to receive education in the official language or languages of their choice in public educational institutions, where it is reasonably practicable. The purposes of these constitutional interventions include a commitment to make sure that there is equity, and it is also an effort towards redressing the injustices of the past. In order to ensure the effective access to, and implementation of this provision, the state must consider all reasonable educational alternatives, including single medium institutions, considering - (a) equity; (b) practicability; and (c) the need to redress the results of past racially discriminatory laws and practices. (The Constitution of South Africa, n.d.)

2.4 The Constitution of South Africa

The South African constitution recognizes the injustices of the past, and thus conferred official status to all historically disadvantaged languages; among other things, it compels the state to take practical and reasonable measures to make sure that the dreams of our people, who were excluded linguistically in the past, are realized. These dreams include the free use of these languages in all public spaces, forceful linguistic development campaigns, and the provision of opportunities and conditions for these languages to not only develop but to flourish. These dreams and aspirations are protected in the Constitution of the Republic of South Africa of 1996. In terms of this supreme law of the country, national government and provincial governments may use any official language for the purposes of government, of course taking into account factors such as usage, practicality, expense, regional circumstances and the balance of the needs and preferences of the population as a whole in the province concerned. The national government and each provincial government must use at least two official languages in official functions. Municipalities must consider the language usage and preferences of their residents. The national government and provincial governments, by legislative and other measures, must regulate and monitor their use of official languages, and all official languages must enjoy parity of esteem and must be treated equally.

The constitution further introduced other measures to emphasize the importance of languages, hence it provided for the establishment of the Pan South African Language Board (PanSALB).

The Constitution of the Republic of South Africa (1996) states that this Board was established by national legislation for the following reasons:

(a) To promote, and create conditions for, the development and use of
   - (i) all official languages;
   - (ii) the Khoi, Nama and San languages;
   - (iii) and sign language;

(b) To promote and ensure respect for
   - (i) all languages commonly used by communities in South Africa, including German, Greek, Gujarati, Hindi, Portuguese, Tamil, Telegu and Urdu; and Arabic, Hebrew, Sanskrit and other languages used for religious purposes in South Africa.

To date, the constitution has provided a platform to allow for the equalization and inclusion of all South African languages, although there are some in our mist who are still cynical about the significance of these interventions. I guess their doubts are influenced by the attitudes and historical role of the English language in the country. Gough (n.d.) explained this as follows:

> English has been both a highly influential language, and a language influenced, in different ways and to different degrees, by processes of adaptation within the country's various communities… amongst the African majority, English has typically been seen as the language of liberation and black unity… and an African language is typically maintained as a solidarity code.

Regardless of how individuals feel about African languages, the Constitution of South Africa must be applauded for promoting equality between South African official languages and for maintaining that each learner should be taught in the language of their choice and insisting that communities should access services in their preferred language.

2.5 History and classification of the Sesotho language

Sesotho is an African language, part of the South-eastern Bantu language family. All nine official African (indigenous) languages of South Africa belong to this family. This language family is part of the Niger-Congo group of languages. The South-eastern Bantu zone is illustrated as follows:

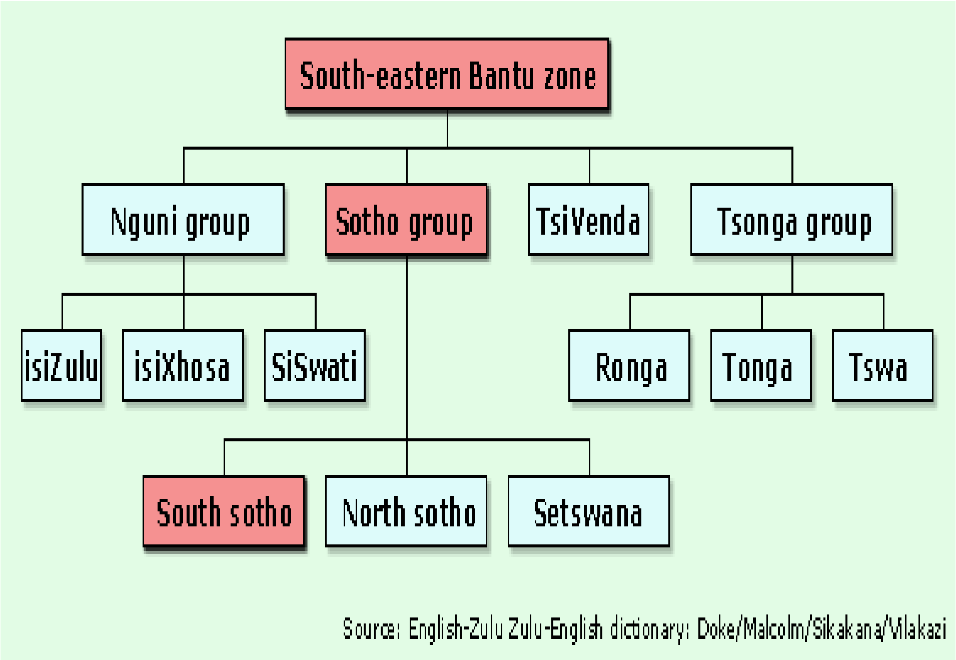


Figure 2.5.1: Structure of South-eastern Bantu zone (https://africanlanguages.com/sesotho/)

The South-eastern Bantu zone consists of four different language groups, namely the Nguni, Sotho, Tshivenda and Xitsonga groups. These different language groups have different language varieties which fall under them and can be distinguished by certain language features that they possess both in written and spoken form. The first group of languages is the Nguni group, which consist of isiZulu, isiXhosa, SiSwati and isiNdebele. Then the second group is Sotho, which consists of Sepedi, Setswana and Sesotho.

Tshivenda is a separate group on its own, whilst the Xitsonga cluster comprises of Ronga, Tonga and Tswa. However, the last two (Tonga and Tswa) are not recognized as official languages in South Africa. Other interventions that have been made by government include the establishment of National Lexicography Units for all the official languages of South Africa.

These National Lexicography Units are dispersed nationally and are based at the institutions of higher learning across South Africa. For example, the University of Free State (n.d.) hosts the *Sesiu sa Sesotho Lexicography Unit*, which is dedicated to the development of Sesotho. Its main purpose is to preserve and record Sesotho by “compiling user-friendly, comprehensive Sesotho dictionaries and other lexicographic products, and to develop and promote Sesotho as a language in all spheres of life”.

The South-eastern Bantu languages of South Africa form part of the Niger-Congo language family and Bantu languages are distinguished from other African languages by means of affixes, which are attached to the root or stem. For example, all nouns of the South-eastern Bantu languages have noun prefixes and are divided according to the noun class systems. Sesotho noun classes are illustrated below in figure 2.5.2.

Class	**Prefix**	**Subject concord**	**Example**	**Translation of example**
1	mo-	o	Mosadi	Woman
2	ba-	ba	Basadi	Women
1a	-	o	Ntate	Father
2a	bo-	ba	Bontate	Fathers
3	mo-	o	Mose	Dress
4	me-	e	Mese	Dresses
5	le-	le	Leleme	Tongue
6	ma-	a	Maleme	Tongues
7	se-	se	Sefate	Tree
8	di-	di	Difate	Trees
9	-	e	Ntja	Dog
10	di-	di	Dintja	Dogs
Classes 11, 12 and 13 are mainly used in languages such as isiZulu and isiXhosa				
14	bo-	bo	Bohobe	Bread
15	ho-	ho	ho tsamaja	to walk
16	-	ho	Fatshe	Down
17	ho-	ho	Hodimo	Up
18	mo-	o	Mose	Abroad
Classes 16, 17 and 18 are called locative classes and they mainly indicate a place or a space.				

Table 2.5.2: Sesotho Noun Classes (http://www.sesotho.web.za/nouns.htm)

2.6 Sesotho dialects, alphabets and vowel variation

Sesotho, which used to be called Sotho or Southern Sotho in South Africa, has different varieties which have certain distinguishing features and are found in different parts of South Africa and Lesotho. Most of the features distinguishing these varieties are found in spoken and written language. In terms of dialects, the following varieties of Sesotho exist: Sekwena, Sephuthi, Setlokwa, and Setaung in the central region; Sekgolokwe in the North-eastern region as well as Serotse. However, the dialects are considered to be mutually intelligible and it is often said that Sesotho has very few dialectal variations. (Austin 2008:101)

Even though Sesotho is closely related to Setswana and Sepedi, the most interesting aspect about it, in contrast to all the Bantu languages, is that it has nine vowels working together at different levels. Some are high, some are low, and others are in the middle (see fig. 2.6.1 below). This is an anomaly in South-eastern Bantu languages. Rialland, Riddouane and Hulst (2015:77), confirms this by stating that:

> One of the world's most complex systems of vowel height is found in the Sotho group of (South-eastern) Bantu Languages spoken in South Africa. These languages have …a thoroughgoing system of partial height assimilation.

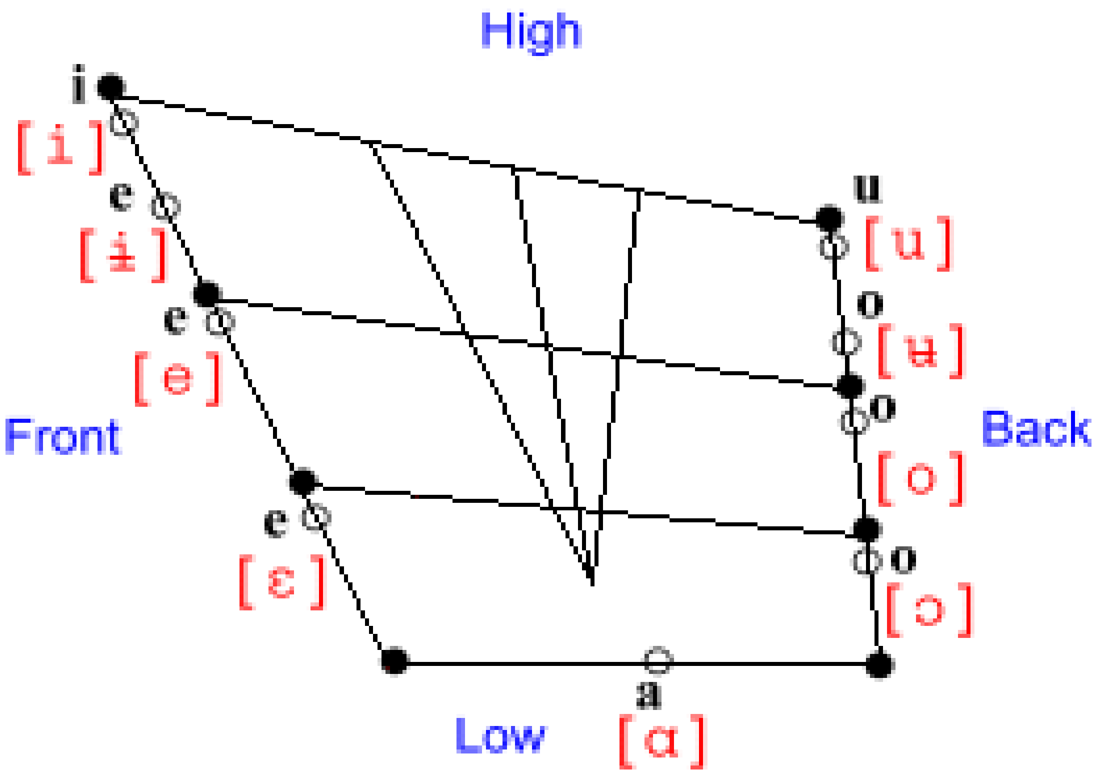
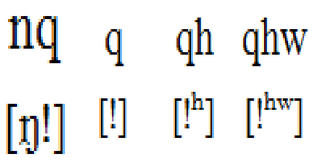


Figure 2.6.1: Sesotho phonetic chart by Doke and Mofokeng (1957)

As opposed to Nguni languages which use a variety of clicks, Sesotho has a minimal number of clicks. In many instances, Sesotho-speaking people find it difficult to pronounce clicks from other languages. Sesotho has the following click variations:

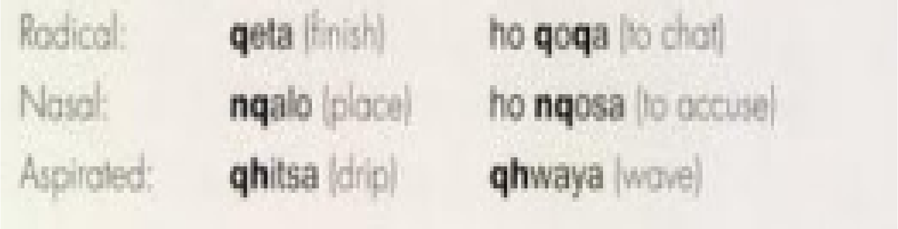


nq	q	qh	qhw
[ŋ!]	[!]	[!^{h}]	[!^{hw}]

Figure 2.6.2: Sesotho clicks (1) (https://www.omniglot.com/writing/sesotho.htm)

According to Austin (2008:101), Sesotho has a large system of forty consonants, and these consonants include click consonants which are borrowed from Khoisan and Nguni languages. However, it can be argued that Sesotho has only one click sound that has different variations: ‘q’ [!] ‘qh’ [!^{h}], qhw [!^{hw}] and ‘nq’ [ŋ!]. In other words, the radical or primary click sound, ‘q’

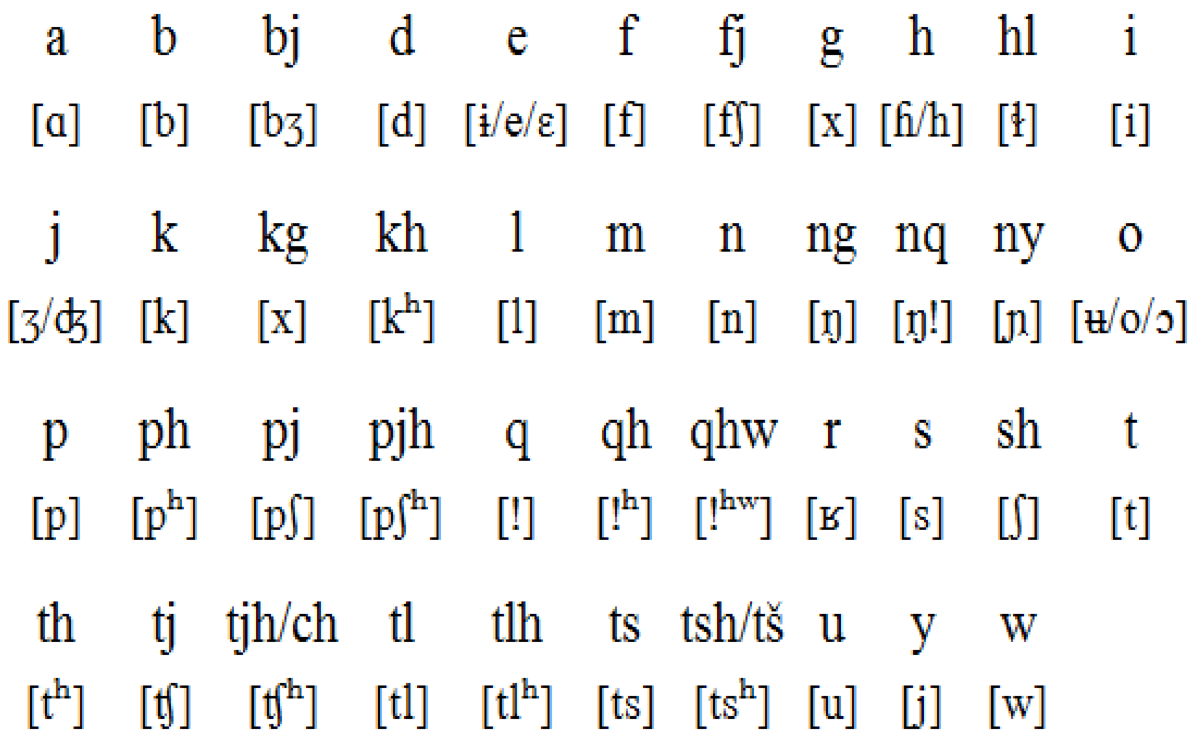
[!] can either be nasalized or aspirated. For example, qh [!^{h}], is aspirated, and ‘nq’ [ŋ!] is a nasalized version of the radical alveolar click, ‘q’ [!].



Radical:	qeta (finish)	ho qoqa (to chat)
Nasal:	nqalo (place)	ho nqosa (to accuse)
Aspirated:	qhitsa (drip)	qhwaya (wave)

Figure 2.6.3: Sesotho clicks (2) by Austin (2008: 101)

Below is a diagram of the Sesotho consonants and vowels together with their phonetic representations. The consonants and vowels found in Sesotho are defined in the International Phonetic Alphabet (International Phonetic Association 1999) and according to McLeod (2010:54-55), each language contains different subsets of the consonants and vowel – with Sesotho having 40 consonants (including clicks).



a	b	bj	d	e	f	fj	g	h	hl	i
[ɑ]	[b]	[bʒ]	[d]	[ɨ/e/ɛ]	[f]	[fʃ]	[x]	[ɦ/h]	[ɬ]	[i]
j	k	kg	kh	l	m	n	ng	nq	ny	o
[ʒ/ʤ]	[k]	[x]	[k^h]	[l]	[m]	[n]	[ŋ]	[ŋ!]	[ɲ]	[ʉ/o/ɔ]
p	ph	pj	pjh	q	qh	qhw	r	s	sh	t
[p]	[p^h]	[pʃ]	[pʃ^h]	[!]	[!^{h}]	[!^{hw}]	[ʁ]	[s]	[ʃ]	[t]
th	tj	tjh/ch	tl	tlh	ts	tsh/tš	u	y	w	
[t^h]	[ʧ]	[ʧ^h]	[tl]	[tl^h]	[ts]	[ts^h]	[u]	[j]	[w]	

Figure 2.6.4: Southern Sotho alphabetic representation and pronunciation (https://www.omniglot.com/writing/sesotho.htm)

2.7 The Sesotho speech community

Sesotho is mostly spoken in South Africa, Lesotho and in some parts of Zambia. It is an official language in South Africa and Lesotho. In Lesotho it is one of the two official languages, which are Sesotho and English, while in South Africa it is one of the eleven official languages. In South Africa a large speech community of Sesotho is found in the Free State province. However, other speech communities of Sesotho are also found in Gauteng and North West and in the northern parts of the Eastern Cape province. There is a rough estimation of about 3.5 million Sesotho-language speakers in South Africa, more than 85% of the Lesotho population speaks Sesotho as their first language, and there are also small Sesotho-speaking communities in Namibia and Zambia. (Accredited Language Services, n.d.).

2.8 Sesotho orthography

Sesotho was one of the first African languages to be reduced to writing and the process was done by Thomas Arbousset, Eugene Casalis and Constant Gosselin, who were French missionaries of the Paris Evangelical Mission who arrived in Lesotho in 1833. The Morija printing press in Lesotho was then established by the trio and the first book to be published by Casalis was *Etudes sur la Langue Sechuana* which was a grammar book published in 1841. Casalis' work was then taken over by the Reverend A. Mabille who later compiled a list of Sesotho words. Major efforts were carried out by the missionaries, and the development of the Sesotho language flourished. In 1872, John Bunyan's *Pilgrim's Progress* was translated into Sesotho and then later a translation of the Bible was completed by the missionaries. (Accredited Languages Services, n.d.)

2.9 The beginning of Sesotho literature

The introduction of writing in Sesotho language paved the way for the development of Sesotho literature. These efforts were also pioneered by the missionaries who wanted to 'enlighten' and convert and, more importantly, introduce Basotho to the Bible.

> The missionaries arrived in Lesotho with the sole purpose of converting Basotho to Christianity. To this end they first had to learn Sesotho. They did this by putting into writing whatever they learnt from their prospective converts. As painstaking

> as it was, their effort laid the first foundation for the Sesotho orthography as we know it today. By the end of 1878 the Bible was available in Sesotho…The missionary work proceeded extremely slowly. In order to facilitate the scriptural teachings, it became necessary to teach prospective converts to read and write. The ability to read and write would enable them to read and sift through the word of God on their own. This led to the establishment of both the Bible School and the Teacher Training School at Morija, in Lesotho… From the aforegoing it is clear that Sesotho literature was not only conceptualised within a Christian context but rather also intended to propagate Christian teachings… Sesotho literature was subjected to censorship. It had to serve the wishes of its producers. (Free State Government, 2003).

Some of the very first works of Sesotho literature include a compilation of Sesotho oral traditions, eminently titled *Mekhoa ea Basotho le maele le litsomo (Customs and Stories of the Sotho*). This was authored by Azariele M. Sekese and published in 1893. However, the most prominent figure in Sesotho literature is Thomas Mokopu Mofolo who wrote a number of novels including, *Moeti oa bochabela* (The Traveller of the East). (Accredited Language Services, n.d.)

2.10 Orthography: Lesotho Sesotho (LSe) vs South African Sesotho (SASe)

The orthographies of Lesotho Sesotho (LSe) and South African Sesotho (SASe) differ in the way sounds are written or joined to form words. However, the pronunciation is generally the same. For example, Sesotho word for ‘sins’: in LSe the word ‘sins’ would be written as *libe* whereas in SASe it would be written as *dibe*. So ‘l’ and ‘d’ differ orthographically.

To emphasize this argument Ngobeni (2010: 249) sums it up as follows:

> Even though most written African Languages use an adaptation of the Latin orthography, most languages spoken across borders have not unified or harmonized their orthographies. For instance, Setswana is spoken in Botswana, South Africa and Namibia; however, while speakers of the language from all these countries can understand each other without any problems in the spoken form, reading each

> other's scripts presents a problem because the language is written with different orthographies in these countries. Another example is Sesotho, which is spoken in South Africa and Lesotho, but uses different orthographies in these two countries. The lack of standardization and unified orthographies makes the sharing of scholarly publications written in indigenous languages difficult and even impossible.

2.11 News and Media

When it comes to broadcast news and media in South Africa, each language has a specific radio station dedicated to it. A radio station that broadcasts in Sesotho and is also dedicated to Sesotho language development is Lesedi FM. In television, however, different languages are grouped together. Please note that the public broadcaster, the SABC (South African Broadcasting Corporation), is the only media institution that has this provision, or rather directive, in its mandate to promote African languages. Private commercial broadcasters are not compelled to accommodate the diverse linguistic communities in our country, because they are driven mainly by the bottom line and other commercial considerations.

South Africa has a long and well-developed print media tradition. However, there are disparities in terms of use and recognition of different languages in this space. *Bona* magazine is the only prominent source of information for Sesotho-speaking communities in print media.

> South Africa currently has the freest press in the continent, with over 20 daily newspapers as well as a slew of strong and independent smaller private and radio stations. (Bradley, Bradley and Fine 2001:54)

African languages have a future as languages of print, but this future is contingent upon, among other things, the adeptness of editors and journalists of African-language print media in forging links with educational and cultural institutions and movements, as well as with other writers in African languages beyond those in the world of media. (Ramadiro, 2013)

During apartheid, radio broadcasting was mainly done in English and Afrikaans, but the post-apartheid era brought some changes, and Ally and Lissoni (2017, par 4) explain these as follows:

> After 1994, as a result of the new reconfigurations which were intended to indicate a turning point away from the separatist ethnic politics propounded by the apartheid government and the realignment with their new values of a unitary South Africa, vernacular stations came up with new names and logos… Radio Sesotho was relaunched to Lesedi FM…

Lesedi FM is a South African radio station, broadcasting semi-nationally and streaming to the world. It was founded on 01 June 1960, as Radio Sesotho, and it focuses on entertaining, informing, educating and empowering South African citizens … [particularly] Sesotho-speaking and Sesotho-understanding people. It is number three amongst the TOP 10 national favourite Radio Stations in South Africa and in Gauteng. The Station's vision is to be a source of enlightenment to its listeners, broadcasting full spectrum, quality programmes to Sesotho-understanding and -speaking communities throughout South Africa. It is head-quartered at the South African Broadcasting Corporation (SABC) in Westdene, Bloemfontein but has studios in Auckland Park, Johannesburg as well. (Lesedi FM, n.d.)

Bona magazine is a South African monthly magazine, reaching over 3 million people every month. It explains itself as South Africa's most read monthly consumer magazine and the go-to brand for today's black family. It is the only South African magazine available in four official languages - English, isiZulu, Sesotho and isiXhosa. At over 60 years old, the brand is one of the country's oldest magazines and this credibility and authority carries through in its content. (Caxton Magazines, n.d.) *Bona* is included in this research because of the role it plays in the development and promotion of Sesotho.

Therefore, *Lesedi FM* and *Bona* (Sesotho edition) are the two most prominent communication platforms for Sesotho-speaking and Sesotho-understanding communities. In addition to being a key information resource, these platforms play a major role in promoting and protecting the Sesotho language in South Africa.

In Lesotho, there is a variety of media outlets. Some of these outlets are owned by the private sector and others are public entities. Since 1998, when the government opened up the media sector to independent media houses, the number of media houses owned by private corporations has grown significantly, although the state electronic media continue to dominate coverage in all areas of the country. Radio Lesotho and Lesotho Television (LTV) are both owned and controlled by the state [and] tend to favour and reflect the government's position on all issues, including electoral matters. Lesotho has approximately 18 newspapers and periodicals, none of which are dailies. The print media consist of various newspapers in the Sesotho language and four English-language weekly newspapers - *The Post*, *The Survivor*, *The Public Eye* and *The Mirror*, which are mostly free from editorial control by the government (Musanhu, 2009)

However, the differences in media coverage in South Africa and Lesotho are worth noting and they show disparities in the number of media used in each country.

2.12 Conclusion

It can be concluded that the African continent has more than 1,500 languages and these are divided into five main language families, namely: Afro-Asiatic, Niger-Congo, Nilo-Saharan, Khoisan and Austronesian. Sesotho falls under the South-eastern Bantu group which is in turn categorized under the Niger-Congo family, and is mostly spoken in South Africa and Lesotho (with small communities in Zambia and Namibia).

Sesotho is one of the official languages of South Africa. Furthermore, it was one of the first African languages to be reduced to writing by French missionaries. Lesotho Sesotho (LSe) and South African Sesotho (SASe) are spoken in the same way, and their speakers understand each other. However, they are written differently, hence it is difficult for academics from either one of these countries to share their knowledge or at least use each other’s literary works.

Lesedi FM and *Bona* magazine have been a source of information and the main tool of communication for Sesotho-speaking and -understanding communities. They play a major role in the promotion and development of Sesotho. They both aim to educate, inspire, and entertain their audiences.

LSe has been advantaged by the fact that it is in competition with only English in Lesotho as they are the two official languages in that country, compared to South Africa where Sesotho is one of eleven official languages. Lesotho has many newspapers written in Sesotho as compared to South Africa, and due to the different orthographic systems used in LSe and SASe, it is often hard for scholars of this language residing in Lesotho to share information with their counterparts in South Africa, and vice versa.

The South African constitution has provided a platform for the promotion and development of all South African languages, and this dissertation responds exactly to that challenge. It also intends to add to the limited research in Sesotho.

Chapter 3: Compilation and utilization of Sesotho corpora

3.1 Introduction

> Worldwide, the compilation, querying and application of electronic corpora has undoubtedly revolutionaised studies as well as descriptions of the structure and use of languages … if scholars of African languges are to take their rightful place in the new milleniun, it is plain that the active compilation, querying and application of electronic corpora should become an absolute priority. (De Schryver and Prinsloo, 2000: 89).

Compiling corpora has been a priority in the quest to develop and compile user-friendly dictionaries. Corpus linguistics has given rise to a more detailed study of language through the use of corpora. Thus, according to O'Keeffe and McCarthy (2010:7):

> …it provides a means for the empirical analysis of language and in so doing adds to its definition and description.

The compilation and utilization of corpora in any language is an important aspect in compiling dictionaries because, once data is compiled, lexicographers are able to use it instantly as a powerful resource to study language. They are able to study the frequency of the occurrence of words in a language, which is then taken into account in deciding the selection versus omission of lemmas in the dictionary. They are also able to study words in context, e.g., through the keywords in the context function of corpus-query programs (also known as concordance lines), to obtain information on different senses, collocations, examples of use, and so on, which in turn helps them to enhance the quality of newly compiled or revised dictionaries. To supplement these points Gouws and Prinsloo (2005:33) point out that:

> On the microstructural level, concordance lines generated by means of corpus-query tools supplement the lexicographer's (native-speaker) intuition.

Before one can delve into the details of the compilation and utilisation of corpora, it is important to first define and understand what a corpus is, and then later to find out how it is compiled using the three stages of the compilation of corpora, namely: corpus design, text collection and text encoding. These will be discussed later.

Different scholars have written about the corpus era and the positive value it has brought, which has led many to undertake research in this field. Many of them realized that corpus studies (or rather the compilation of corpora) is important in any study of language because a corpus documents words/tokens used in a particular language and is also able to identify a lot of factors which otherwise could not have been identified if the corpus was not used. These factors range from grammar studies, phonetics and written and oral communication, to lexicographic issues such as lemma selection, treatment of lemmas in dictionaries, etc.

> Unsurprisingly, corpora have been used extensively in nearly all branches of linguistics, including, for example, lexicographic and lexical studies, grammatical studies, language variation studies, contrastive and translation studies, diachronic studies, semantics, pragmatics, stylistics, sociolinguistics, discourse analysis, forensic linguistics and language pedagogy. (McEnery and Xiao 2010:364)

Corpora set new ways of dealing with languages and also made simple, easy and manageable, the work of language practitioners. At present, it is recommended that every language practitioner needs to have a corpus in order to study their language effectively and accurately. Hence Jeffery (2000:7l) believes that “it has become widely accepted that a well-designed corpus is a prerequisite for study of any language”. This idea justifies the notion that a corpus is a useful tool in compiling user-friendly dictionaries with fewer errors.

> A corpus can either be small or large depending on its usage and what the compiler hopes to achieve at the end of their study. However, it is generally perceived that the bigger the corpus, the better simply because a large one will ensure that more conclusions are drawn and provide more evidence on a specific issue. Consider the following quote and statistics given by the corpus of the Brigham Young University. Corpus size is incredibly important, in terms of the richness of the corpus data. A tiny one-million-word corpus is extremely limited in terms of the

> phenomena that it can study - compared to a 400-million-word corpus, where there might be 400 times as much data. (Davies, n.d.)

However, the situation is quite different for African languages because they have limited resources and the following quote better illustrates this limitation:

> "Big corpus" is a relative term. For lesser-resourced languages with a limited number of printed materials such as many of the African languages, a corpus of 10 million words can be regarded as a "big corpus". (Prinsloo, 2015:286)

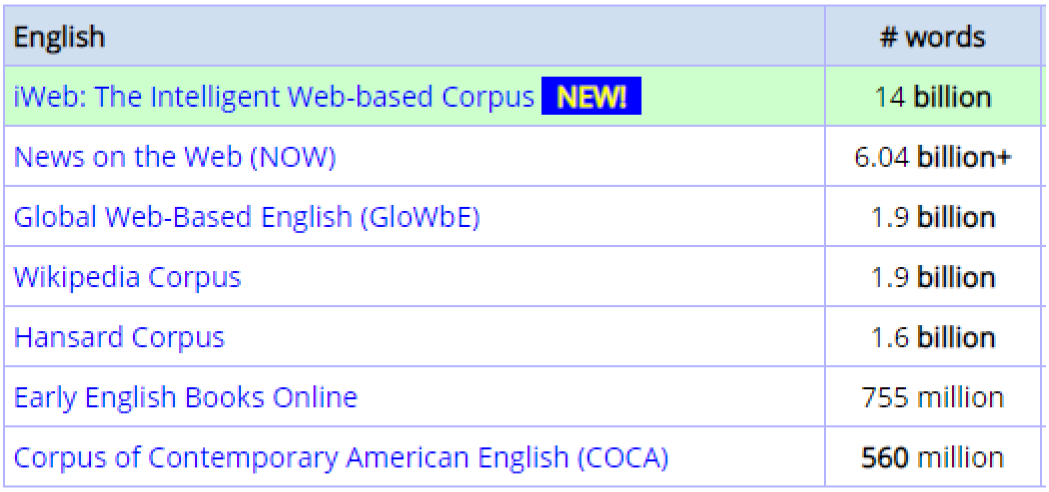
In CCURL (2014) the reasons for this limitation are explained as follows:

> 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.

It is important to also note that for one to compile a corpus, they need to establish the kind of corpus they wish to develop which could range between the following types: Type 1) full-text, sample and monitor; Type 2) synchronic and diachronic; Type 3) general and terminological; Type 4) monolingual, bilingual and plurilingual; Type 5) language corpus; Type 6) single, parallel-2, parallel-3…; Type 7) central and shell; Type 8) core periphery. (Atkins, Clear and Ostler 1991:13)

3.2 History of electronic corpora

Electronic corpora date back to the 1960s when the first corpus was established. It contained about one million words referred to as tokens. It is known as the Brown University Standard Corpus of Present-day American English which is also known as the Brown Corpus. (Francis and Kučera, 1964). The Brown Corpus was purely used for linguistic research. Then later in the 1980s, the corpus era dawned upon lexicography and was spearheaded by Collins COBUILD in Birmingham which published the first corpus-based dictionary in 1987, namely, the Collins COBUILD Dictionary. Since the establishment of Collins COBUILD Dictionary, corpus sizes rose markedly to the extent that by the end of 1987 the COBUILD RESERVE Corpus had about 13 million words (Renouf, 1987:7-10). Today, corpus sizes for the major languages of the world run into billions of tokens.



English	# words
iWeb: The Intelligent Web-based Corpus NEW!	14 billion
News on the Web (NOW)	6.04 billion+
Global Web-Based English (GloWbE)	1.9 billion
Wikipedia Corpus	1.9 billion
Hansard Corpus	1.6 billion
Early English Books Online	755 million
Corpus of Contemporary American English (COCA)	560 million

Table 3.1: English corpora (https://corpus.byu.edu/)

The Longman Lancaster English Language corpus was then developed and consisted of 30 million words (Leech 1992:3-4). It was part of a commercial project named the Longman Corpus Network (LCN) and was also one of the three major corpora developed under LCN. The British National Corpus then followed between the years 1991 and 1995. It consisted of a 100 million words (Leech, 1992:1). The Bank of English was then developed at the University of Birmingham and consisted of over 320 million words by 1998 (Hartmann and James 1998:12).

3.3 History of electronic corpora: an illustration for African languages

Even though electronic corpora have been compiled in western countries and scholars have diligently contributed to its development, the same cannot be said for African languages – electronic corpora for these are still in the infancy stage and need a lot of attention from African-language scholars. It is for this reason that Prinsloo and De Schryver (2000:89) suggest that, if African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of new corpora should become an absolute priority.

African-language practitioners need to understand that corpora simplify the work of lexicographers and linguists because it enables them to study language in context. Hence Gouws and Prinsloo (2005:8) state that:

> The electronic medium has become increasingly important for the transfer of knowledge, and lexicography had to respond to this. The electronic medium allows lexicographers a wholly new approach to the dictionaries without the space restrictions and limitations which macrostructural ordering and the access structure put on printed dictionaries.

For African languages, it is important to note that corpora are being built for them, specifically for the South African indigenous languages, even though they do not have the necessary funds, corpus traditions, human resources, demand, frameworks, etc. Prinsloo and De Schryver (2000:96) say:

> It is beyond doubt that any first approach to corpora for the African languages cannot even come close either to the size or thoroughness that characterizes today's major English corpora. Nor do we have the necessary funds, nor the necessary demand to name but a few, to warrant such tremendous efforts.

3.4 Corpus size

The quotation above has also touched on another subject which also affects the process of corpus compilation – namely the issue of size. To analyse this issue, we have to ask the following questions: 1) Does the size of the corpus really matter? 2) Does bigger really mean better?

3.4.1 Does the size of a corpus really matter?

Since a corpus is a body of text, written or spoken, any text with 100 or 1,000 000 tokens is still regarded as one – because what matters is the fact that there is a group of tokens which has been collected and in which some kind of language analysis can take place. Although this is the case, a relatively large corpus is preferable because a large corpus has high possibilities of including a large number of words/terms used in a specific language, which will give a better understanding of the language. Taking into consideration the level of corpora development in African indigenous languages, the issue of size is a tricky one because these languages are allocated few resources and lag behind because they do not have a large body of written material that can be used for corpora studies. However, spoken corpora will form the basis of corpora development in these previously disadvantaged languages, and Gouws and Prinsloo (2005:21-22) explain:

> The study of oral data can pinpoint words which tend to be used more frequently in oral versus written communication. Unfortunately, most corpora around the world lack sufficient data from spoken sources. The reason for this is that there are many logistical problems and ethical factors involved in the collection of spoken data.

With regard to the size of corpora, many scholars followed the Brown Corpus (Brown University Standard Corpus of Present-Day American English) as a blueprint, since it is one of the earliest major electronic corpora which has roughly one million words (Francis and Kučera, 1994). From there, the size of corpora escalated with Birmingham University International Language Database (COBUILD) with the COBUILD main corpus containing about 7.3 million words in 1982; and 13 million words were assembled in the COBUILD RESERVE Corpus by 1987 (Renouf 1987:7-10). Even though the Longman Corpus Network (which consisted of three other major corpora) was still being built, one of its branches (the Longman Lancaster English Language Corpus) had about 30 million words (Summers, 1993:184-201). Another corpus which contained about 100 million words, also compiled between 1991 and 1995, was the British National Corpus (BNC) (Kennedy, 1998:50). The Bank of English, initiated in 1991 at the University of Birmingham, had 320 million words by 1998 (Hartmann and James, 1998:12)

3.4.2 Does bigger really mean better?

A large corpus essentially means that many sources have been consulted (which takes us back to the issue of balance versus representativeness). However, bigger usually means better because a wealth of sources have been consulted.

Another approach to be noted is the creation of organic corpora, which entails adding all available material to the corpus over a period of time. In this regard, the organic corpus must be allowed to grow and live so that it adequately represents the language. Gouws and Prinsloo (2005:25) note that "an interesting approach to the compilation of corpora, and one that fits the situation for African languages like a glove, is the concept of organic corpora …"

Organic corpus creation will be discussed in more detail below.

> From the information provided, it seems easy for one to therefore conclude that size could matter, as the big corpora that have been built since the beginning of the corpus era have set a bar which a corpus of any language needs to reach. However, the African languages have a totally different story to tell because "… a neatly designed collection is not possible and the whole selection process eventually boils down to all available texts for the specific language" (Gouws and Prinsloo 2005:24).

3.5 Corpus designs

Designing a corpus entails finding the information the compiler wants to focus on and its overall purpose; however, for their corpus to represent the 'entire' language, it is advisable to draw information from a variety of domains. Considering figures 3.5.1, 3.5.2 and 3.5.3, it was important for the compilers to understand the kind of data they wanted to extract from each corpus – hence they had to have different genres of a particular language. "A corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully-constructed." (Sinclair, 2004).

Text categories		**Number of samples in each category.**
Brown corpus		LOB corpus
Press: reportage.	44	44
Press: editorial	27	27
Press: reviews	17	17
Religion	17	17
Skills, trades and hobbies	36	38
Popular lore	48	44
Belles-lettres, biography, essay	75	77
Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
Learned and scientific writings	80	80
General fiction	29	29
Mystery and detective fiction	24	24
Science fiction	6	6
Adventure and westerns	29	29
Romance and love story	29	29
Humour	9	9
TOTAL	**500**	**500**

Table 3.5.1: Basic composition of Brown and LOB corpora (Johansson and Hofland, 1989:2)

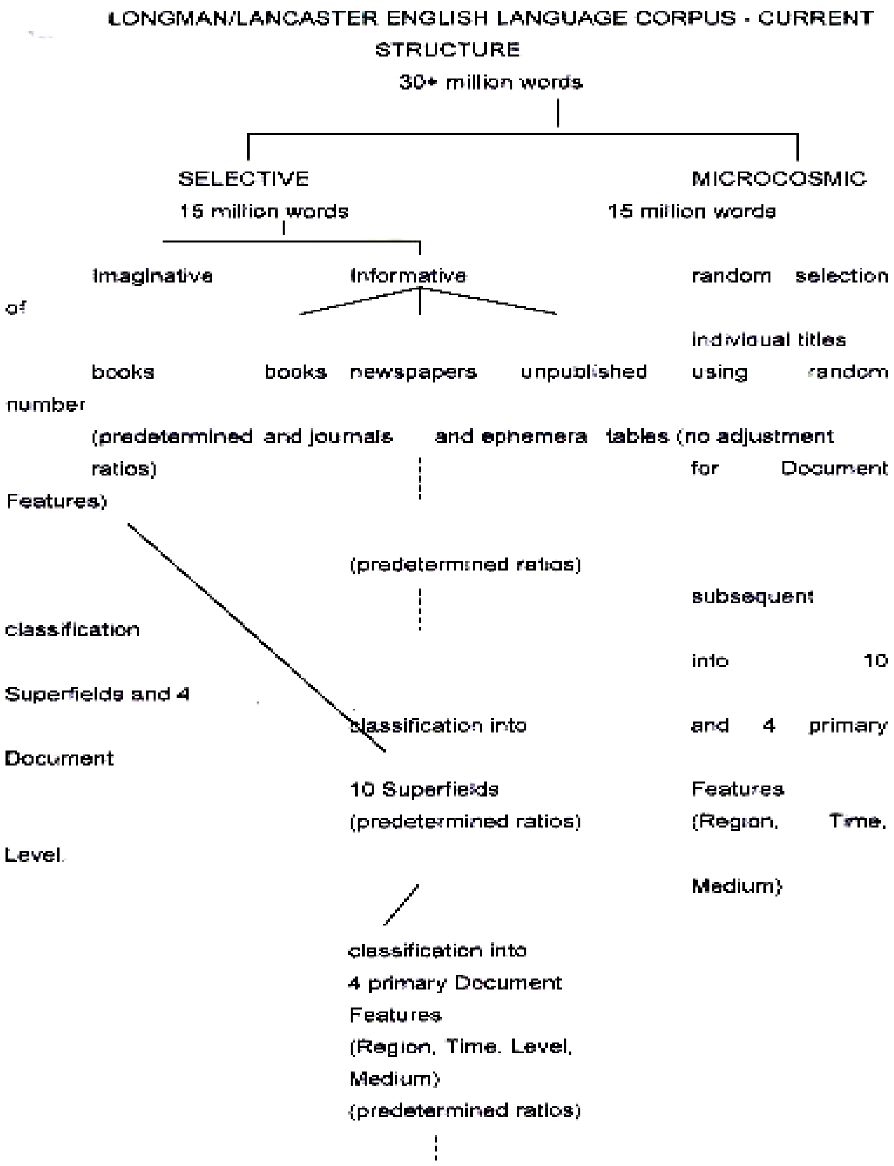
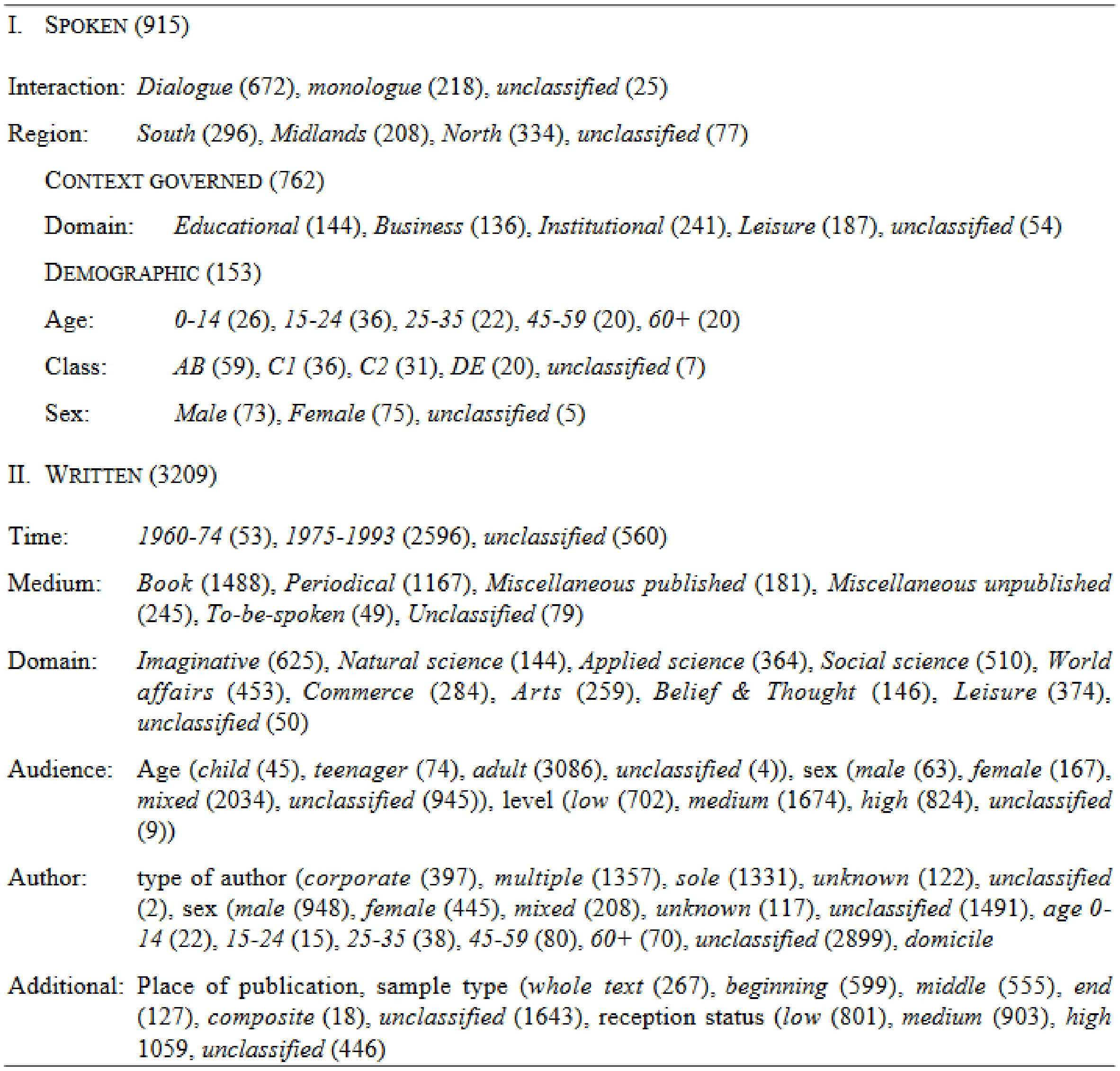

LONGMAN/LANCASTER ENGLISH LANGUAGE CORPUS - CURRENT STRUCTURE
30+ million words
SELECTIVE
15 million words
MICROCOSMIC
15 million words
Imaginative
Informative
random selection of individual titles using random number tables (no adjustment for Document Features)
books (predetermined ratios)
books and journals
newspapers and ephemera
unpublished
(predetermined ratios)
classification into 10 Superfields (predetermined ratios)
classification into 4 primary Document Features (Region, Time, Level, Medium) (predetermined ratios)
classification Superfields and 4 Document Level
subsequent into 10 and 4 primary Features (Region, Time, Medium)

Figure 3.5.2: Longman/Lancaster English language corpus – current structure by Summers (1993: 201)



I. SPOKEN (915)

Interaction: *Dialogue* (672), *monologue* (218), *unclassified* (25)

Region: *South* (296), *Midlands* (208), *North* (334), *unclassified* (77)

CONTEXT GOVERNED (762)

Domain: *Educational* (144), *Business* (136), *Institutional* (241), *Leisure* (187), *unclassified* (54)

DEMOGRAPHIC (153)

Age: *0-14* (26), *15-24* (36), *25-35* (22), *45-59* (20), *60+* (20)

Class: *AB* (59), *C1* (36), *C2* (31), *DE* (20), *unclassified* (7)

Sex: *Male* (73), *Female* (75), *unclassified* (5)

II. WRITTEN (3209)

Time: *1960-74* (53), *1975-1993* (2596), *unclassified* (560)

Medium: *Book* (1488), *Periodical* (1167), *Miscellaneous published* (181), *Miscellaneous unpublished* (245), *To-be-spoken* (49), *Unclassified* (79)

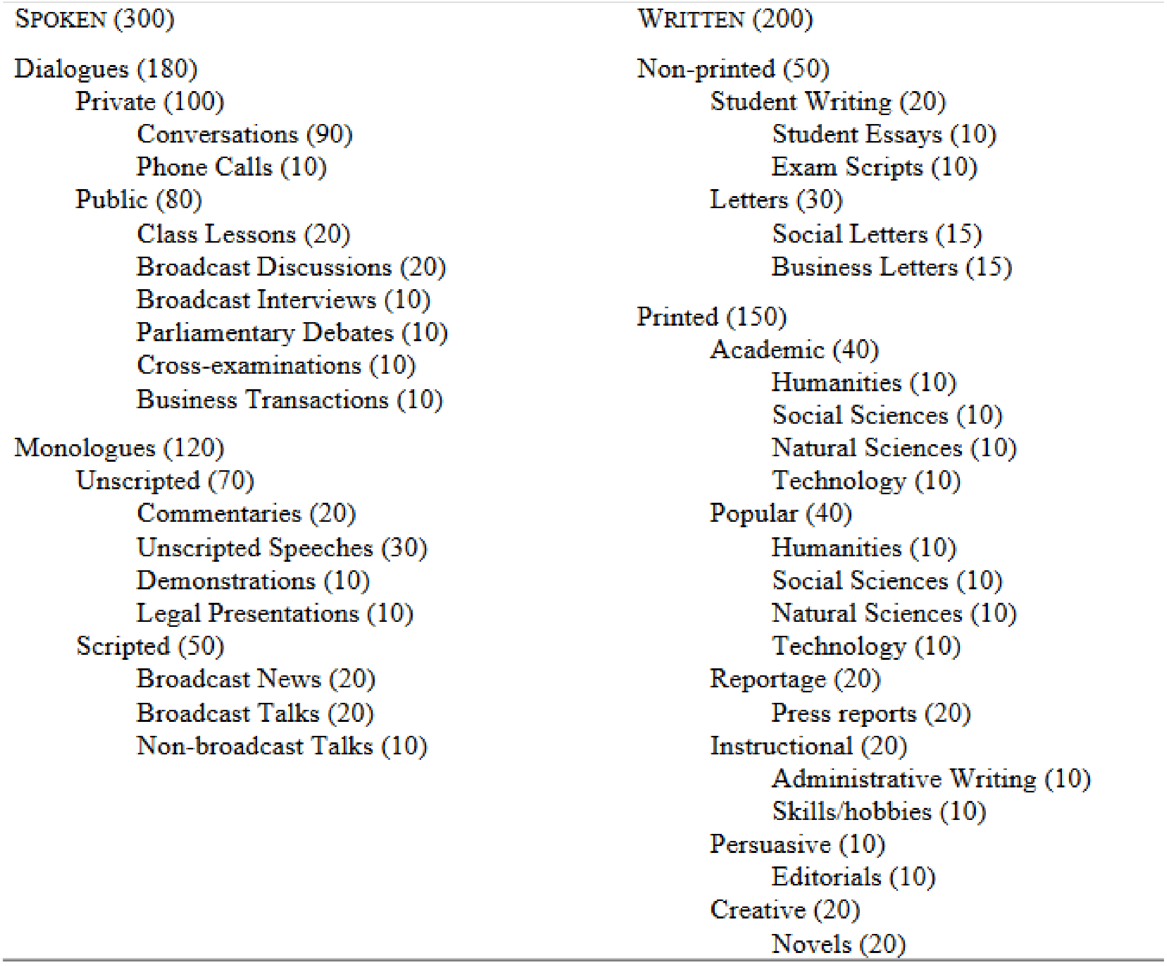
Domain: *Imaginative* (625), *Natural science* (144), *Applied science* (364), *Social science* (510), *World affairs* (453), *Commerce* (284), *Arts* (259), *Belief & Thought* (146), *Leisure* (374), *unclassified* (50)

Audience: Age (*child* (45), *teenager* (74), *adult* (3086), *unclassified* (4)), sex (*male* (63), *female* (167), *mixed* (2034), *unclassified* (945)), level (*low* (702), *medium* (1674), *high* (824), *unclassified* (9))

Author: type of author (*corporate* (397), *multiple* (1357), *sole* (1331), *unknown* (122), *unclassified* (2), sex (*male* (948), *female* (445), *mixed* (208), *unknown* (117), *unclassified* (1491), *age 0-14* (22), *15-24* (15), *25-35* (38), *45-59* (80), *60+* (70), *unclassified* (2899), *domicile*

Additional: Place of publication, sample type (*whole text* (267), *beginning* (599), *middle* (555), *end* (127), *composite* (18), *unclassified* (1643), reception status (*low* (801), *medium* (903), *high* 1059, *unclassified* (446)

Figure 3.5.3: The British National Corpus (https://www.dbthueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00010791/corp_design.pdf)



SPOKEN (300)

- Dialogues (180)
    - Private (100)
        - Conversations (90)
        - Phone Calls (10)
    - Public (80)
        - Class Lessons (20)
        - Broadcast Discussions (20)
        - Broadcast Interviews (10)
        - Parliamentary Debates (10)
        - Cross-examinations (10)
        - Business Transactions (10)
- Monologues (120)
    - Unscripted (70)
        - Commentaries (20)
        - Unscripted Speeches (30)
        - Demonstrations (10)
        - Legal Presentations (10)
    - Scripted (50)
        - Broadcast News (20)
        - Broadcast Talks (20)
        - Non-broadcast Talks (10)

WRITTEN (200)

- Non-printed (50)
    - Student Writing (20)
        - Student Essays (10)
        - Exam Scripts (10)
    - Letters (30)
        - Social Letters (15)
        - Business Letters (15)
- Printed (150)
    - Academic (40)
        - Humanities (10)
        - Social Sciences (10)
        - Natural Sciences (10)
        - Technology (10)
    - Popular (40)
        - Humanities (10)
        - Social Sciences (10)
        - Natural Sciences (10)
        - Technology (10)
    - Reportage (20)
        - Press reports (20)
    - Instructional (20)
        - Administrative Writing (10)
        - Skills/hobbies (10)
    - Persuasive (10)
        - Editorials (10)
    - Creative (20)
        - Novels (20)

Figure 3.5.4: The International Corpus of English (https://www.dbthueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00010791/corp_design.pdf)

3.6 Balanced versus representative corpora

Ideally, a corpus should be balanced, and representative of a language. When the question of representativeness is put forward, it important to understand that it does not occur on its own, rather it is accompanied by balance and sampling. Hence McEnery, Xiao and Tono (2006:125) state that a "representative corpus is achieved by balancing the corpus through sampling a wide range of text categories which are defined primarily in terms of external criteria. Gouws and Prinsloo (2005:17) also state that: "No modern dictionary can be representative if it is not based on a reliable corpus"

However, developing a representative corpus is always a challenge because one has to take into account various factors as described below:

> There are two major problems when trying to compile a representative corpus. First, it has to be determined what dimensions of language variation should be represented. Second, it has to be determined whether sampling should be based on what is produced or on what is perceived. For example, does a representative corpus of sports writing include samples from all sports writers proportionally to how much they have written, or proportionally to how many people read their articles? (Stefanowitsch 2003).

In most instances, linguists agree to disagree whenever the notion of balanced and representative corpora is discussed, for the reason that many believe that corpora need to be balanced in such a way that data is extracted from all sources and from all the genres across a spectrum of spoken or written data. When they talk of a balanced corpus, they essentially mean that the corpus compiler will be visiting varieties of data from different genres. As a result, Kennedy (1998:20) contends that:

> A general corpus is typically designed to be balanced, by containing texts from different genres and domains of use including spoken and written, private and public, formal and informal, etc.

Kennedy (1998:52) also states that:

> For a corpus to be representative there must be a clearly analysed and defined population to take the sample from. This therefore means that once a set of samples is extracted from different genres, it is safe to say that one's corpus represents all aspects of a language because each and every genre has been visited and a sample extracted from each. In this way, regardless of the size of the corpus all tokens will be represented in one way or the other. However, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre.

Another point illustrated by Summers (1993:186) is that "we believe that unless the corpus is representative, it is *ipso facto* unreliable as a means of acquiring lexical knowledge."

Linguists agree to disagree when the notion of balanced or representative corpora is debated, for the reason put forward by Kennedy (1998:62). “Questions associated with ‘representative’ and ‘balance’ are complex and often intractable”.

Another important aspect is illustrated by Prinsloo (2000:92) when he maintains that:

> It seems as if a corpus will never be balanced because there are too many parameters, and it seems as if a corpus will never be truly representative of all language usage, either, as it is impossible to determine the population.

On the same note, the issue of representativeness is illustrated in the following text, with specific reference to COBUILD:

> COBUILD has always insisted that it is impossible to create a corpus that is truly representative of the language, and have focused on size of corpus rather than balance. (Kilgarriff, 1997:150).

3.7 The organic corpus

The notion of an organic corpus was conceptualized and introduced by Atkins, Clear and Ostler (1992:1, 4, 6), who explained it as follows:

> A corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing, living language. [...] In order to approach a "balanced" corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weaknesses identified and reported. In the light of this experience and feedback, the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually. [...] In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus - however "unbalanced" - to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts.

The above quote illustrates that a corpus compiler needs to find a way of modelling and structuring the corpora so that they are balanced and represent the language. An organic corpus goes through a lot of growth and many changes because language evolves over time, and so the corpus has to reflect that change and be in a better position to give lexicographers good results (in terms of omission and inclusion, examples of use, collocations, etc.). It is for this reason that De Schryver and Prinsloo (2000:92) maintain that:

> The minimum requirement for any organic corpus is thus that the corpus compiler(s) will have attempted to put some structure in assembling the range of electronic texts. Within this framework, any first attempt at compiling an organic corpus will at least result in a structured corpus.

Corpus compilation for Sesotho got wings to reach new heights with the launch of the South African Digital Languages Resources (SADiLaR). In terms of its self-description, it is a national centre supported by the Department of Science and Technology (DST) as part of the new South African Research Infrastructure Roadmap (SARIR), which aims to provide research infrastructure across the public research system, building on existing capabilities and strengths, and drawing on future needs. (SADiLaR, n.d.)

SADiLaR has an enabling function, with a focus on all official languages of South Africa, supporting research and development in the domains of language technologies and language-related studies in the humanities and social sciences [and] supports the creation, management and distribution of digital language resources, as well as applicable software, which are freely available for research purposes through the Language Resource Catalogue. (SADiLaR, n.d.)

Large-scale corpus creation is envisaged, and publishers thus far are quite willing to waive copyright on sources that could be included. This great initiative will bring good change in research for previously disadvantaged languages and is a great step towards raising the status of these languages.

3.8 The design of the Sesotho corpus

> The optimal design of a corpus is highly dependent on the purpose for which it is intended to be used. (Kennedy 1998:70)

Drawing on the information provided above (on the different types of texts), this study will focus on the first option, namely the Full Text for lexicographic application. For this, a collection of material in the Sesotho language, i.e., books (novels, fiction and non-fiction, poetry, drama, etc.), magazines and newspapers will be used to extract data. The corpus will be consulted to obtain answers on frequency of use, paraphrase of meaning, selection of translation equivalents, sense distinction, examples of usage, collocations, etc. Since a lot of work is done by the University of Pretoria and the Pretoria Sepedi Corpus is well researched and annotated, it would be advisable for the Sesotho Corpus to follow in its footsteps. Hence De Schryver and Prinsloo (2000:96) point out that:

> …even though size and thoroughness of the South African corpora so far cannot compare with those of the English corpora, the point is that they cannot be compiled in isolation since important aspects from other comparable corpus projects must be taken into consideration.

The research is based on text material rather than spoken data. Prinsloo is of the opinion (one-on-one communication) that it is approximately ten times more expensive, labour intensive and time consuming to compile an oral corpus compared to a text corpus of the same size. So it follows that the little work done thus far in the compilation of Sesotho corpora was focused on written texts. De Schryver and Prinsloo (2000:94) name three basic ways of building a corpus by entering the texts into the computer. This process involves downloading a variety of well-selected documents from the internet or retrieving texts already on computer disk, (re)keyboarding, (i.e., typing) of handwritten documents or even printed matter into computer files, and scanning of printed matter by means of OCR (Optical Character Recognition) software such as *OmniPage*, which can be accessed through the following websites, among others:

1) *OmniPage*, (https://www.nuance.com/print-capture-and-pdf-solutions/optical-character-recognition/omnipage.html)
2) ABBYY, (https://www.abbyy.com/en-eu/finereader/)
3) Readiris, (https://www.irislink.com/EN-US/c1729/Readiris-17--the-PDF-and-OCR-solution-for-Windows-.aspx)

Electronic transfer of existing electronic documents, especially containing text without errors, is by far the quickest and most effective way to build a corpus – in contrast to scanning or (re)typing of sources, which is time consuming.

According to De Schryver and Prinsloo (2000:95), text encoding has the following sub-categories: a) word tokenisation, b) part of speech tagging, c) lemmatisation, d) syntactic parsing, and e) mark up. Word tokenisation involves segmenting words into free-standing words while part of speech tagging (b) focuses on assigning word classes on all words to show their belonging grammatically by way of part of speech hence the term POS-tagging. Lemmatisation deals with finding lemmas of words whilst syntactic parsing focuses on analysing sentences. Finally, mark-up refers to the electronically annotated texts which have been processed using the initial four stages.

3.9 Conclusion

This chapter dwelled on the history of the electronic corpus, highlighting the fact that the pioneers of corpora did a sterling job with very limited resources. The foundation they laid, however, was very strong and we continue to exploit it to this day; that is why, for instance, the Brown Corpus is still referred to as the original landmark of corpus creation. Although electronic corpora have been around for decades and various languages have taken advantage of its opportunities and conveniences, the progress of African languages in this regard remains lamentable.

The issue of the size of corpus was also discussed extensively in this chapter. The debate of whether size matters, or even whether a big corpus means a better corpus, is an interesting one. As a result, there is no consensus among scholars on the matter, but it is generally assumed that "bigger is better". In this chapter it is also emphasized that, in order to create a perfect corpus design, one would need to consult a wide range of sources, include data from a variety of genres of a particular language, and so on. Another interesting debate that is covered at length in this chapter is the level of importance regarding balance and representativeness in corpora. Prinsloo (2000:92) answers this perfectly: “It seems as if a corpus will never be balanced because there are too many parameters, and it seems as if a corpus will never be truly representative of all language usage, either, as it is impossible to determine the population”.

The history of the organic-corpus concept was discussed as well, highlighting the fact that, like languages, organic corpora also change and evolve. It is emphasized that the design of a Sesotho corpus should follow the standard conventions, and that Sesotho lexicographers should use as references, the languages that have rich lexicographic heritages. A desperate call has also been made to Sesotho scholars and authorities to make the compilation of Sesotho corpora a priority.

Chapter 4: Corpus-enabled Macrostructural compilation

4.1 Introduction

Dictionaries have evolved over the years and have been revolutionised with the use of various methods which lexicographers use to solve language or lexicographic problems they are faced with. A corpus is the main component of such methodologies and is used more and more by scholars who realise its important role in modern-day language studies. Before the advent of computers and computer-based lexicography, the macrostructure was compiled manually, and this method is time consuming for lexicographers. However, everything changed for the better when the first electronic corpus was created in the 1980s and the first electronic corpus-based dictionary was compiled. This meant that most, if not all, the lexicographer’s problems would be solved by this great invention, which revolutionised the study and practice of lexicography. The study of languages changed for the better as linguists were able to find specific answers to specific problems which could otherwise not have been found in the absence of electronic corpora. The advent of corpora also meant that the solutions would be found faster, thereby saving lexicographers time and helping them to determine frequency counts, concordances, etc.

The macrostructure of a dictionary is a fundamental aspect to consider during dictionary compilation, as it forms the core or backbone of the dictionary. It can be explained simply as the compilation of the different components of a dictionary, i.e., front matter, back matter and central text. Most of the lexicographer’s labour goes into the central text. The work that is done in this component includes the compilation of the lemma list and subsequent treatment of those lemmas. On the macrostructural level, the focus is on the selection and presentation of lemmas.

Consider the following definitions of the macrostructure of a dictionary:

According to Van Sterkenburg (2003:157):

> The macrostructure refers to the way in which entries are arranged within a dictionary.

Hüllen (2006:179) explains the macrostructure by noting that it refers to:

> …the dictionary as a whole; more precisely to the alphabet of its sections.

While Bergenholtz and Tarp (1995:15) define the macrostructure in the following way:

> Macrostructure is the lexicographic term used to describe the arrangement of the stock of lemmata in the wordlist. A dictionary may have one or more macrostructures, according to the number of wordlists. Macrostructure can be systematic i.e. arranged to a systematic classification or alphabetic e.g. strictly alphabetic arrangement and arrangement according to the nest principle.

It is vital to note that these scholars come to the same conclusion that it has something to do with the lemma sign list, the sequence of lemmas and their ordering, as well as what to include and exclude in the dictionary. This in essence means that any lexicographer undertaking a dictionary project of some sort must be aware and acknowledge the roles played by these macrostructural factors.

However, bearing in mind the definitions above, a slightly similar but different approach is taken by Nielsen (1994:74-76) who states that:

> The lexicographic macrostructure deals with more than the arrangement of lemmata. It may be appropriate to compare the word macrostructure with the word macrocosm. In the same way as the term macrocosm applies to the whole universe, the macrostructure of a dictionary may be regarded as parallel to macrocosm. Thus, it may be said that the lexicographic macrostructure applies to the dictionary as a whole and not merely the arrangement of lemmata. In other words, the macrostructure of a dictionary may be described as the organisational structure of the dictionary which is concerned with the sequential relationship between the macrostructural component.

The explanation by Nielsen (1994) is more informative and inclusive. In other words, it expands the definition so that it covers some of the most important aspects of the macrostructure. It touches on the originality and the core of any macrostructure of any dictionary and sums up areas of focus which need to be looked at by lexicographers. This definition has more substance because it does not only look at the arrangement of entries, the number of entries or their organisation. Rather it also delves into the relationship of those macrostructural components as well and emphasizes the consideration of the dictionary as a whole. It further acknowledges the relationships that coexist within words, and also explains that words do not occur in isolation, they are, instead, a part of a complex system.

4.2 Types of Macrostructures

A dictionary is expected to adhere to certain rules and follow a particular structure. However, it is also important to note that as the dictionary is compiled, it takes its own form and shape. It is influenced by factors such as the audience, data extracted from corpora, interpretation of results, etc. Thus, rules should not be solid or rigid, because each dictionary is different from the next and should be allowed to take on a life of its own. In as much as compilers construct dictionary structures in a specific way (see figure 4.2.1), these rules should be allowed to change as the dictionary takes on a life of its own (Frawley, Hill, Munro 2002:153).

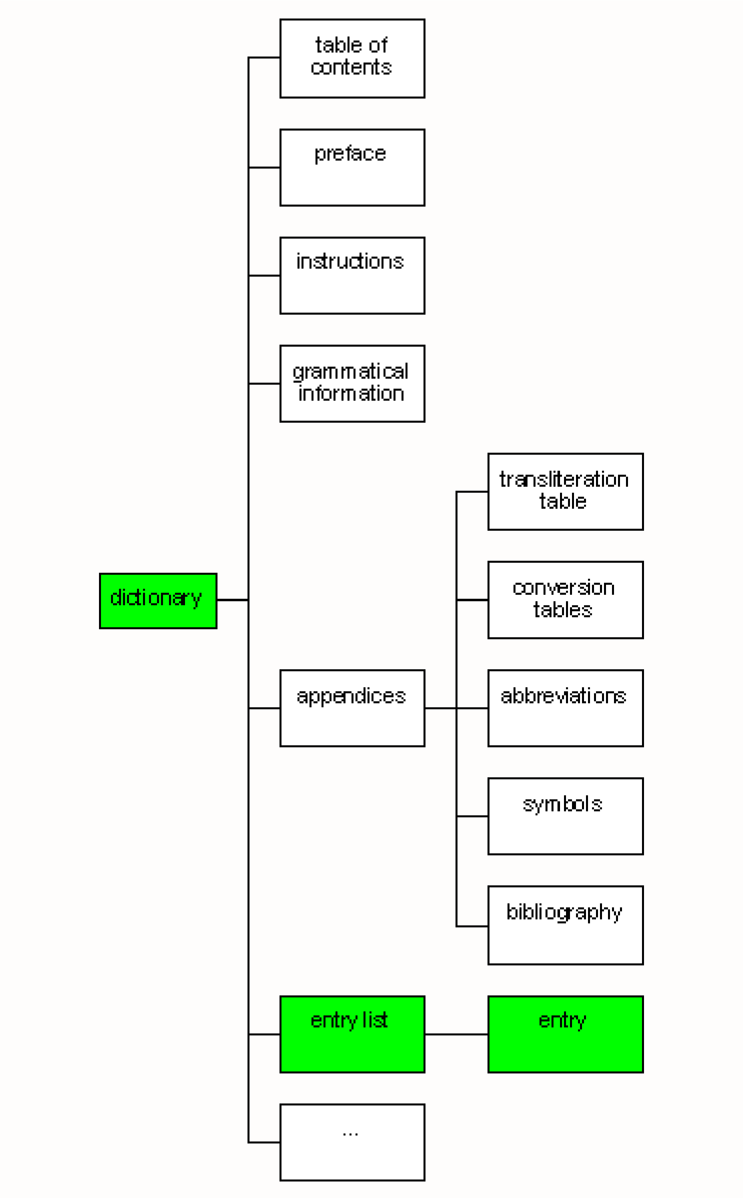


Figure 4.2.1: The structure of a dictionary (https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/dict_structure.html)

In an ideal world, all dictionaries must have, as a prerequisite, fully functional macrostructures with all the necessary information. However, it is also possible for dictionaries to have underdeveloped macrostructures. Hence Nielsen (1994:78, 84) states that:

> Ideally, every dictionary should have a complex macrostructure which should apply to (at least) a list of contents, a preface, a user's guide and the (alphabetical)

> wordlist. Dividing the dictionary into the separate macrostructural components will facilitate the user's search for relevant information in connection with the use of the dictionary as well as in connection with the interpretation of the information contained in it.

There are at least two types of macrostructures: the simple macrostructure and the complex macrostructure (see figure 4.2.2).

> A simple macrostructure may be described as a lexicographic macrostructure which applies to only two macrostructural components… [and] is typically made up of a preface and an alphabetically arranged wordlist…The basic reason for limiting the maximum number of macrostructural components to two is that it simplifies identifying the interrelationship between only two separate components. (Nielson 1994:77)

The simple macrostructure reflects a relationship between two different macrostructural components, i.e., between the list of contents and preface, or between user guide and list of abbreviations, and so on, and would typically include a normal preface together with a list of lemmata. However, should one or more components be added to the two that already exist, another type of macrostructure will surface. This type of macrostructure is referred to as a complex macrostructure and contains enough information to help the user understand the dictionary and be able to use it better (Nielson 1994: 77-78). It also has all the components included in simple macrostructures together with the introduction to the dictionary, which mainly consists of the entry structure, core sense and sub-senses, labels used, grammar, spelling and inflection and – most importantly – the pronunciation of lemmas dealt with in the dictionary. (Ortlepp, 2007:3)

The richness or lack of information on these types of macrostructures affect the user of the dictionary. If the information contained in the dictionary is richer, it will benefit the user in many ways. It has been emphasized that every dictionary has to be compiled with the user in mind. In other words, a well-compiled dictionary should not confuse, mislead or frustrate the users or simply fail to provide the information that the target user might look for. The following illustration is a diagram of the different macrostructures. Nielson (1994:78) goes on to state that "the main problem with a macrostructure like this is that all the information intended to guide the user in the search for information is contained, or rather concealed in the preface". In this argument, the terms 'contained' and 'concealed' are already an indication that the user will be left frustrated in an age where lexicographers are urged to compile user-friendly dictionaries. A question that will have to be answered would be "why must information intended for users be hidden?" But then again, we must remember that lexicographers have their reasons for taking some routes and ignoring the others. Hence, Nielson (1994:78) saw that "it does seem, however, that only few dictionaries have simple macrostructures since many contain more than two macrostructural components".

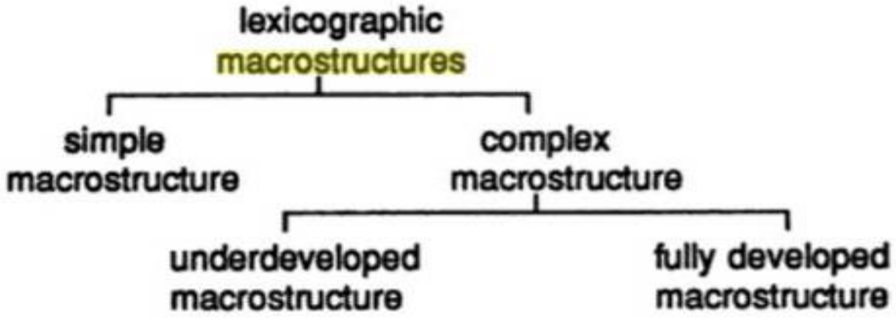

lexicographic macrostructures
simple macrostructure
complex macrostructure
underdeveloped macrostructure
fully developed macrostructure

Figure 4.2.2: Different macrostructures by Nielsen (1994:85)

4.3 The macrostructural components

Concerning macrostructural components, the focus here is on the front matter, central text and back matter. The central text includes activities such as word-frequency counts, compilation of the lemma sign list, and so forth.

> When trying to describe this macrostructural interrelation, it may be advantageous to divide the dictionary as a whole into three parts. The first part of this trichotomy is the front matter or fore matter … which may be described as the part of the entire dictionary which immediately precedes the wordlist. (Nielsen 1994:86)

Nielsen (1994) makes the following remarks concerning the front matter of a dictionary. The first phenomena (the front matter) refers to all the information included in the dictionary that comes before the lemma list. This information includes the list of contents, which gives the user an idea of what the dictionary contains, followed by the preface which entails the author's explanatory remarks, then comes the user guide with steps on how to use the specific dictionary including how information can be obtained. Then next is the list of abbreviations, the table of contents and, in some cases, the field introduction (also see figure 4.3.1). The order may vary from dictionary to dictionary and, to some extent, the macrostructural components included in dictionaries may differ according to the purpose, specificity or generalisation of the dictionary and/or the target user and age (e.g., learner or adult).

> According to its purpose and target group, a dictionary may include an almost infinite variety of separate macrostructural components in its front matter. (Nielsen 1994:105)

Below is an example of how a front matter of a dictionary is structured.

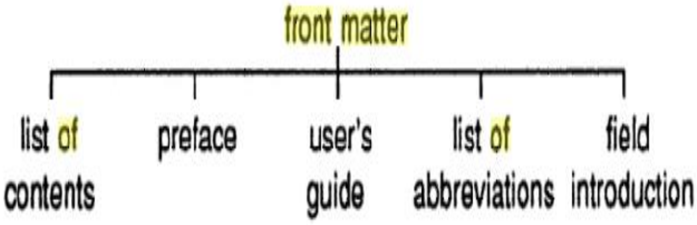


Figure 4.3.1: Presentation of the front matter by Nielsen (1994:105)

The other feature is the back matter, which includes all the information after the word list. This information normally consists of an appendix, cross-references and in some cases other works of the author(s) or compiler(s). Hence Gouws et al, (2013:59) suggest that "the ordered set of text compound constituents that follows the wordlist constitutes the dictionary back matter … outer text that belong to the back matter are also known as back matter texts".

The central text is the other element of Nielsen's trichotomy, and is the most important component of a dictionary. The idea is to check the number of times words or lemmas occur in a variety of texts – in different genres if the corpus being compiled is for general purposes or a specific genre if the corpus is for specific purposes. Hence Gouws and Prinsloo (2005:30) point out that "publishers are normally very specific and prescriptive regarding the number of lemmas to be treated."

Restriction on the size of the dictionary plays a major role in printed dictionaries as they are always constrained by space limitations and if, for example, a publisher requires that a dictionary must have 500 pages with 20,000 lemmas, it is the responsibility of the lexicographer to adhere to such rules. The challenge therefore rests on the lexicographers to make sure that they lemmatise words most likely to be looked for by their target users, and this means using their own intuition and extracting such words from the corpus with special attention to frequency of occurrence. This means that they do not have to rely solely on their intuition for lemma selection.

4.4 Corpora and the Macrostructure

Since the age of compiling dictionaries using corpora dawned, upon modern dictionary compilation the use of corpora proved to be a success in more ways than one.

> Corpora are especially useful in that they make it possible to separate the frequent and average from the once-offs; to distinguish the typical from the oddities. (De Schryver 2003:29).

Corpora are discussed here because they are the first task in any modern lexicographic project and building one allows for a full macrostructure to be derived from them. Hence De Schryver (2003:11) avers that "the most important stage in any modern lexicographic project is to build an electronic corpus where a full macrostructure of a dictionary can be derived from".

African-language dictionaries are still presumed to be in their infancy. For Sesotho dictionaries in particular to take their rightful place in the new millennium, its lexicographers need to start using modern tools to study language, and to compile well-structured, lexicographically planned, corpus-based and user-friendly dictionaries. It is for this reason that De Schryver and Prinsloo (2000:291) state that "the challenge is thus to compile dictionaries of African languages according to the latest trends and modern approaches in lexicography". There is a huge demand and need for African-language dictionaries to be of a high standard and of high lexicographic quality. Even though these standards may not be met in a single day, it is crucial that African-languages scholars start taking advantage of the benefits of modern-day electronic corpora and corpus query tools in order to achieve these goals.

To compile a macrostructure that is of a high standard, it is important to use the latest trends and approaches in lexicography. This means compiling and using corpora as the basis for all lexicographic activities. The use of corpora in the compilation process ensures that frequency lists, selection strategies (on what to exclude and what to include in the dictionary) as well as how to order the entries are on point in a user-friendly way.

4.5 Frequency tests

The frequency tests are done by taking different bodies of texts from different genres, finding out how many times each word occurs in each text, and then drawing conclusions from the results. Of significance for the compiler of a general dictionary are words that occur frequently across a variety of sources in written texts or oral communication. Depending on the target users and the nature and size of the dictionary, decisions can be made on what to include and what to omit from the dictionary. African languages, and Sesotho in particular, are resource-scarce languages which do not have a lot of written material and have a limited oral corpus.

It is for this reason that Gouws and Prinsloo (2005:30) illustrate that "on the macrostructural level, word frequency counts are an extremely useful tool in the compilation of a lemma list for new dictionaries". This is also true for a dictionary that is being revised because, instead of having to compile a new lemma list, a lexicographer can make use of existing data and then arrange it well by taking the results from the frequency tests and then comparing them to the existing dictionary. In this way the lexicographer will have a clear indication of which words to lemmatise and which ones to omit.

Gouws and Prinsloo (2005:30) also go on to say that:

> If the required number of lemmas is say 10,000, the lexicographer could for instance isolate the top 10,000 types from a frequency list, lemmatise it and supplement the lemma list with lower frequency items from the frequency list until the desired number has been met.

By using frequency lists, the lexicographer avoids the shortcomings of dictionary compilation done solely on intuition, i.e., not omitting words likely to be looked for by the target users.

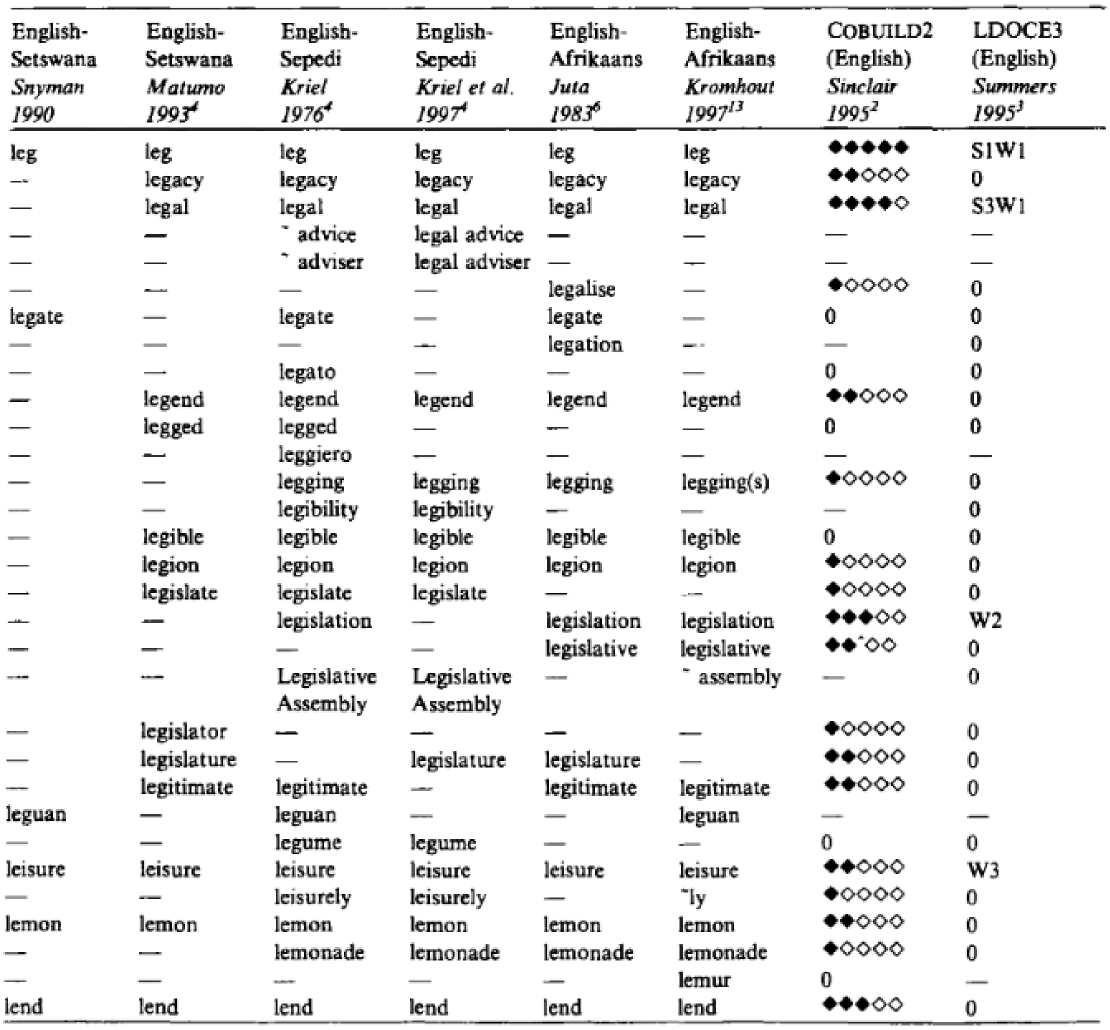
> The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compilation of a dictionary. This can take place simply because the lexicographer had not encountered such words (Snyman et al. 1990: preface)

4.6 Putting frequencies in the dictionary

Putting frequencies in a dictionary is another step in dictionary compilation which affects how the dictionary is presented. It gives the lexicographer a stance on the kind of material that should be lemmatised based on data extracted from corpora. The higher the frequency of a word, the higher the chances of it being treated in the dictionary. Indicating frequencies in a dictionary does not entail the inclusion of raw frequency counts of words but the use of so-called frequency bands, e.g. the top 5,000 frequencies. This is done by conventions such as filled/shaded diamonds, a star-rating, and symbols such as “W” for frequency in written language “S” for frequency in spoken language. The two final columns in figure 4.6 show two different strategies.

According to De Schryver and Prinsloo (2000:294) the shaded diamonds represent frequency of use, see figure 4.6.

> Five filled diamonds indicate that the lemma sign occurs within the 700 most frequently used words in English, four that it occurs within the first 1,900, three that it occurs within the first 3,400, two that it occurs within the first 6,600 and one that it occurs within the first 14,700.



English-Setswana *Snyman* *1990*	English-Setswana *Matumo* *1993*[4]	English-Sepedi *Kriel* *1976*[4]	English-Sepedi *Kriel et al.* *1997*[4]	English-Afrikaans *Juta* *1983*[6]	English-Afrikaans *Kromhout* *1997*[13]	COBUILD2 (English) *Sinclair* *1995*[2]	LDOCE3 (English) *Summers* *1995*[3]
leg	leg	leg	leg	leg	leg	◆◆◆◆◆	S1W1
—	legacy	legacy	legacy	legacy	legacy	◆◆◇◇◇	0
—	legal	legal	legal	legal	legal	◆◆◆◆◇	S3W1
—	—	˜ advice	legal advice	—	—	—	—
—	—	˜ adviser	legal adviser	—	—	—	—
—	—	—	—	legalise	—	◆◇◇◇◇	0
legate	—	legate	—	legate	—	0	0
—	—	—	—	legation	—	—	0
—	—	legato	—	—	—	0	0
—	legend	legend	legend	legend	legend	◆◆◇◇◇	0
—	legged	legged	—	—	—	0	0
—	—	leggiero	—	—	—	—	—
—	—	legging	legging	legging	legging(s)	◆◇◇◇◇	0
—	—	legibility	legibility	—	—	—	0
—	legible	legible	legible	legible	legible	0	0
—	legion	legion	legion	legion	legion	◆◇◇◇◇	0
—	legislate	legislate	legislate	—	—	◆◇◇◇◇	0
—	—	legislation	—	legislation	legislation	◆◆◆◇◇	W2
—	—	—	—	legislative	legislative	◆◆ˆ◇◇	0
—	—	Legislative Assembly	Legislative Assembly	—	˜ assembly	—	0
—	legislator	—	—	—	—	◆◇◇◇◇	0
—	legislature	—	legislature	legislature	—	◆◆◇◇◇	0
—	legitimate	legitimate	—	legitimate	legitimate	◆◆◇◇◇	0
leguan	—	leguan	—	—	leguan	—	—
—	—	legume	legume	—	—	0	0
leisure	leisure	leisure	leisure	leisure	leisure	◆◆◇◇◇	W3
—	—	leisurely	leisurely	—	˜ly	◆◇◇◇◇	0
lemon	lemon	lemon	lemon	lemon	lemon	◆◆◇◇◇	0
—	—	lemonade	lemonade	lemonade	lemonade	◆◇◇◇◇	0
—	—	—	—	—	lemur	0	—
lend	lend	lend	lend	lend	lend	◆◆◆◇◇	0

Figure 4.6: Diamond representation of frequencies (De Schryver and Prinsloo 2000:294)

4.7 Lemmatisation

According to Gouws and Prinsloo (2005:67), lemmatisation can be defined in an over-simplified way as the selection of a specific form from a given paradigm to be used in a dictionary as the starting point for information retrieval. So, for example, an English lexicographer would select *buy* as the lemma to represent the paradigm *buy*, *buys*, *buying*, etc., and the Sesotho lexicographer can select *bua* to represent the paradigm *bua*, *buile*, *buuwa*, *builwe*, etc. Lemmatisation in African languages in general, and in Sesotho in particular, occurs in two traditions, i.e., the word tradition and the stem tradition. The former is used for disjunctively written languages which consist of Sesotho, Sepedi, Setswana, Tshivenda and

Xitsonga while the latter refers to Nguni languages, namely isiZulu, Siswati, isiXhosa and isiNdebele. Van Wyk (1995:83) points out that “the word tradition originated in the Sotho languages, Venda and Tsonga, and requires that lexical entries be based on complete written words. In the stem tradition which is characteristic of Nguni languages, but is also followed in some Sotho dictionaries, the stem of written words forms the basis of lexical entries”. The stem versus word traditions will be discussed in more detail below.

The Nguni languages follow a conjunctive orthographic system and the Sotho languages a disjunctive way of writing. The differences in the two writing systems are purely on orthographic grounds. Consider the following example:

English	Sesotho	isiZulu
The children are eating bread	*Bana ba ja bohobe*	*Abantwana badla isinkwa*

In the Sesotho phrase, words have been separated whereas in Nguni (isiZulu) they are conjoined. In terms of lemmatisation, the word tradition for the disjunctive languages would lemmatise each of those four words separately, i.e., *bana*, *ba*, *ja* and *bohobe* however, the stem tradition would go as far as lemmatising only the stems of the words, i.e., *-ntwana*, *-dla* and -*nkwa*. With this said, it is important to always remember that different languages as well as different dictionaries require different lemmatisation approaches, strategies and lexicographic traditions. There is no ‘one size fits all’ approach where one method is applicable to all dictionaries and to all languages. As a result, for a lexicographer to have clear lines on all these issues, while at the same time having the target user in mind, it is important to draw on the usefulness of the corpus together with its corpus query tools.

4.8 Lemmatisation approaches, strategies and lexicographic traditions

The lemmatisation approaches, strategies and lexicographic traditions briefly stated above are important aspects during the compilation of the macrostructure, especially for African languages, as the lexicographer has to “negotiate a complex interplay and overlap between (a) lemmatisation approaches (b) lemmatisation strategies, (c) lexicographical traditions, (d) nominal and verbal structures and (e) conjunctiveness versus disjunctiveness” (Gouws and Prinsloo 2005 :68).

> The Bantu-language lexicographer not only has to deal with all of these aspects, but he or she also has to consider the complex interplay within (a) to (d) for each dictionary to be compiled in order to fulfil the needs of the respective target users. Prinsloo (2009:152)

With regards to lemmatisation approaches for African languages, such as Sesotho, lexicographers have successfully utilized five approaches for lemmatising nouns which according to Prinsloo and De Schryver (1998) are in the following order: lemmatising noun stems, lemmatising both singular and plural noun forms, lemmatising only singular noun forms, and lemmatising nouns on the first or third letter. The first approach has proved to be popular amongst lexicographers and many of them consider it as the only systematic and scientific way of lemmatising nouns for African languages. Hence, Guthrie (1971:358) contends that the essential word structure of Bantu languages requires that for indexing purposes the stem of a noun and not its concord be positioned in an alphabetic order, while Ziervogel and Mokgokong (1975:87) suggest that "it is the only scientific method".

The traditional approach, irrespective of whether it is applied in word or stem lemmatisation, entails a situation where lexicographers add words as they encounter them, without any plan on how to lemmatise. The use of this approach means that there are high chances of omission of words which are most likely to be looked for because the lexicographer did not encounter them.

> [This approach is] characteristic of revisions of bilingual dictionaries bridging Sesotho sa Leboa and English or Afrikaans. With each new revision one could see how more words were merely added to these dictionaries. This approach represents the worst situation where the compiler does not employ any selection strategy and even seems to be unaware of the problem of what to include in and what to omit from the dictionary. (Gouws and Prinsloo 2005:71)

The rule-orientated approach deliberately limits lemmatisation of words, especially the treatment of word derivations. This approach utilizes rules which guide the user during information retrieval when applied correctly. This approach is not user-friendly because it entails that a user must know certain rules in order to access information.

> However, any rule-orientated approach runs into serious difficulty with regard to practicality and user-friendliness. (Gouws and Prinsloo 2005:75)

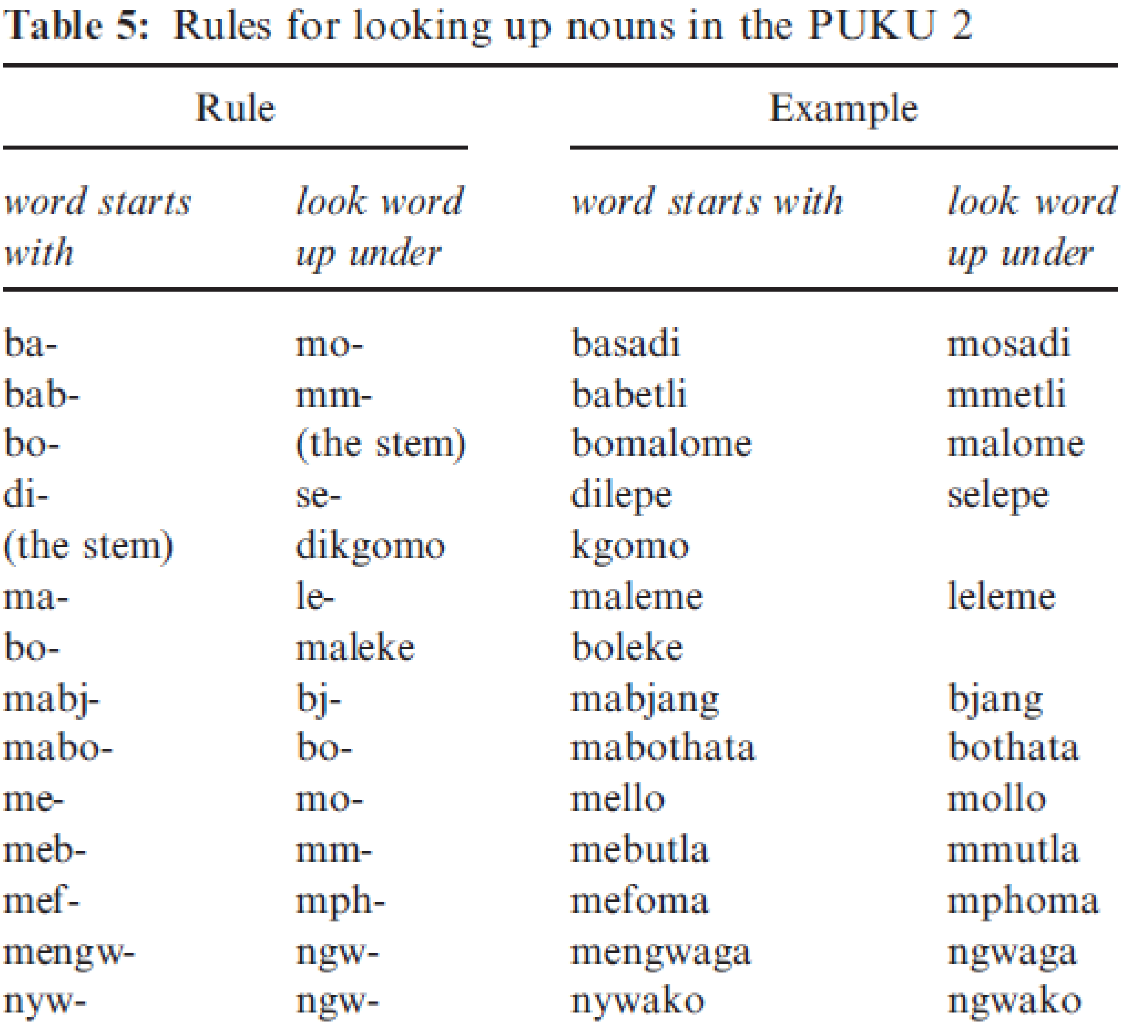


Table 5: Rules for looking up nouns in the PUKU 2

Rule		Example	
word starts with	*look word up under*	*word starts with*	*look word up under*
ba-	mo-	basadi	mosadi
bab-	mm-	babetli	mmetli
bo-	(the stem)	bomalome	malome
di-	se-	dilepe	selepe
(the stem)	dikgomo	kgomo	
ma-	le-	maleme	leleme
bo-	maleke	boleke	
mabj-	bj-	mabjang	bjang
mabo-	bo-	mabothata	bothata
me-	mo-	mello	mollo
meb-	mm-	mebutla	mmutla
mef-	mph-	mefoma	mphoma
mengw-	ngw-	mengwaga	ngwaga
nyw-	ngw-	nywako	ngwako

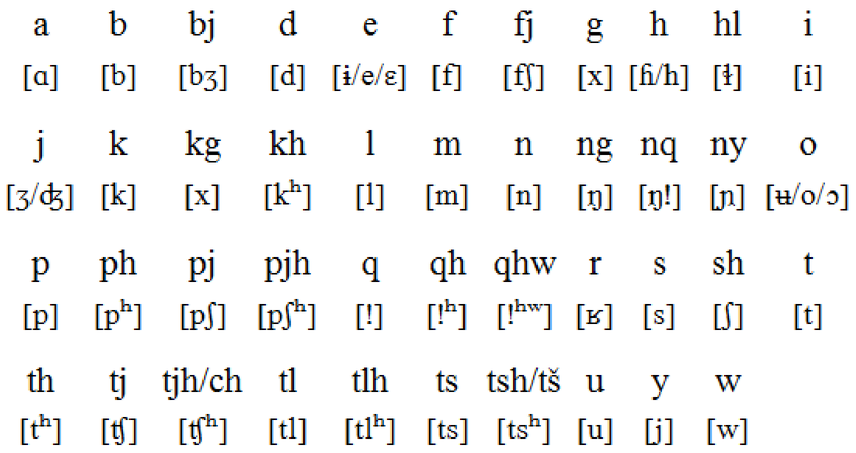
Figure 4.8.1: Example of rule-orientated approach (Prinsloo 2015:160)

The paradigm approach entails a situation where the lexicographer includes all possible derivations of a word either as lemmas or as sub-lemmas. The problem with this approach is the fact that even non-existing words could be included in the dictionary, thereby consuming space (in printed dictionaries) with words that will not be searched for. Omission of important words is also evident here at the expense of words most likely to be looked for.

> [There is an] attempt to enter all nominal and verbal derivations to such an extent that mother-tongue speakers doubt whether many of these derivations are actually and actively used. (Gouws and Prinsloo 2005: 72)

Orthography of the language plays a major role in lemmatisation and has major implications. For disjunctively written languages, such as languages in the Sotho group (Sepedi, Setswana, and Sesotho) including Tshivenda and Xitsonga, lemmatisation is non-problematic and the ratio of token/orthographic word versus lemma is almost 1:1. However, for the conjunctively written languages like isiZulu, isiNdebele, isiXhosa and Siswati, complex lemmatisation processes to isolate stems, affixes and concords are required. In most cases orthography has a direct bearing on lexicographic traditions in Bantu lexicography. (Prinsloo 2015:155)

Alphabetical ordering focuses on the first letter of the stem. This way the user can arrive at the required lemma with ease. This process influences the ordering of the articles as well and is user-friendly because it demands less from the user. This type of ordering prevails in modern-day dictionaries and is the preferred procedure in general descriptive and bilingual dictionaries. However, it is important to note that this access alphabet is not necessarily identical with the standard alphabet as it is used in its everyday sense. (Gouws and Prinsloo 2005:97)



a	b	bj	d	e	f	fj	g	h	hl	i
[ɑ]	[b]	[bʒ]	[d]	[ɨ/e/ɛ]	[f]	[fʃ]	[x]	[ɦ/h]	[ɬ]	[i]
j	k	kg	kh	l	m	n	ng	nq	ny	o
[ʒ/ʤ]	[k]	[x]	[k^{h}]	[l]	[m]	[n]	[ŋ]	[ŋ!]	[ɲ]	[ʉ/o/ɔ]
p	ph	pj	pjh	q	qh	qhw	r	s	sh	t
[p]	[p^{h}]	[pʃ]	[pʃ^{h}]	[!]	[!^{h}]	[!^{hw}]	[ʁ]	[s]	[ʃ]	[t]
th	tj	tjh/ch	tl	tlh	ts	tsh/tš	u	y	w	
[t^{h}]	[ʧ]	[ʧ^{h}]	[tl]	[tl^{h}]	[ts]	[ts^{h}]	[u]	[j]	[w]	

Figure 4.8.2: Southern Sotho pronunciation https://www.omniglot.com/writing/sesotho.htm

Dictionaries that follow a phonetic ordering, assemble their articles phonetically and according to how they are pronounced. According to Gouws et al (2013: 798):

> The order within the wordlist is not based on the written alphabet due to the opinion that spoken language, opposite to written language, is an independent domain whose structure should be analysed based on its own principles.

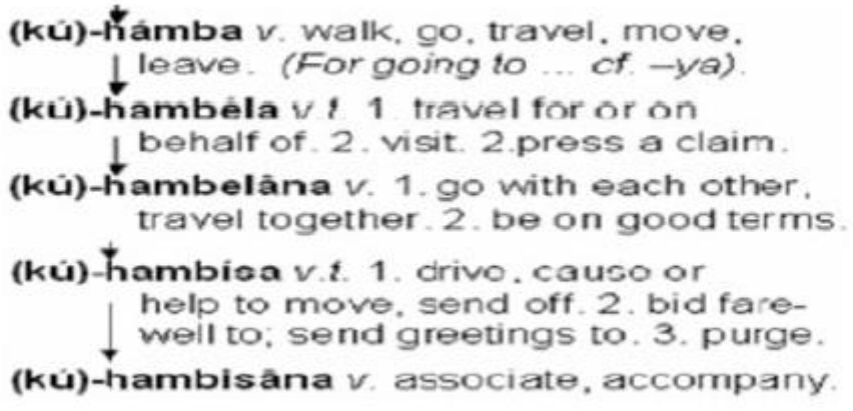
Another point is that this type of ordering puts users under unnecessary confusion and strain because they are expected to know the phonetical rules of a language initially for them to utilize such dictionaries.

> That way, an ordering is established which is closer related to phonetics and phonology, but this order is not a widely known one and untrained users will face problems locating the expressions they are looking for. (Gouws et al 2013: 798)

Left-expanded stem lemmatisation of nouns entails lemmatisation of the full noun with its prefixes but the alphabetical ordering runs on the stem as in figure 4.8.3. Prinsloo (2011:188) suggests that:-

> Lemmatising stems with their prefixes merely added on (left-expanded) is a better option, because the user has the advantage of seeing the full form of infinitive verbs and the full forms of nouns with additional information, such as tonal indication. This strategy is more user-friendly, but stem identification remains problematic and a substantial amount of knowledge of morpho-phonetics is still presupposed.

Left-expanded stem lemmatisation for verbs as described by Gouws and Prinsloo (2005) is the lemmatisation of the verb stem with the infinitive prefix, for example, *kuhamba* ‘to walk’ in Siswati. The alphabetical ordering runs on the first letter of the stem with the infinitive prefix left expanded as for -*hamba* and its derivations in Rycroft’s Concise SiSwati Dictionary (CSD) in (3).



(kú)-hámba *v.* walk, go, travel, move, leave. *(For going to ... cf. –ya).*

(kú)-hambêla *v.t.* 1. travel for or on behalf of. 2. visit. 2.press a claim.

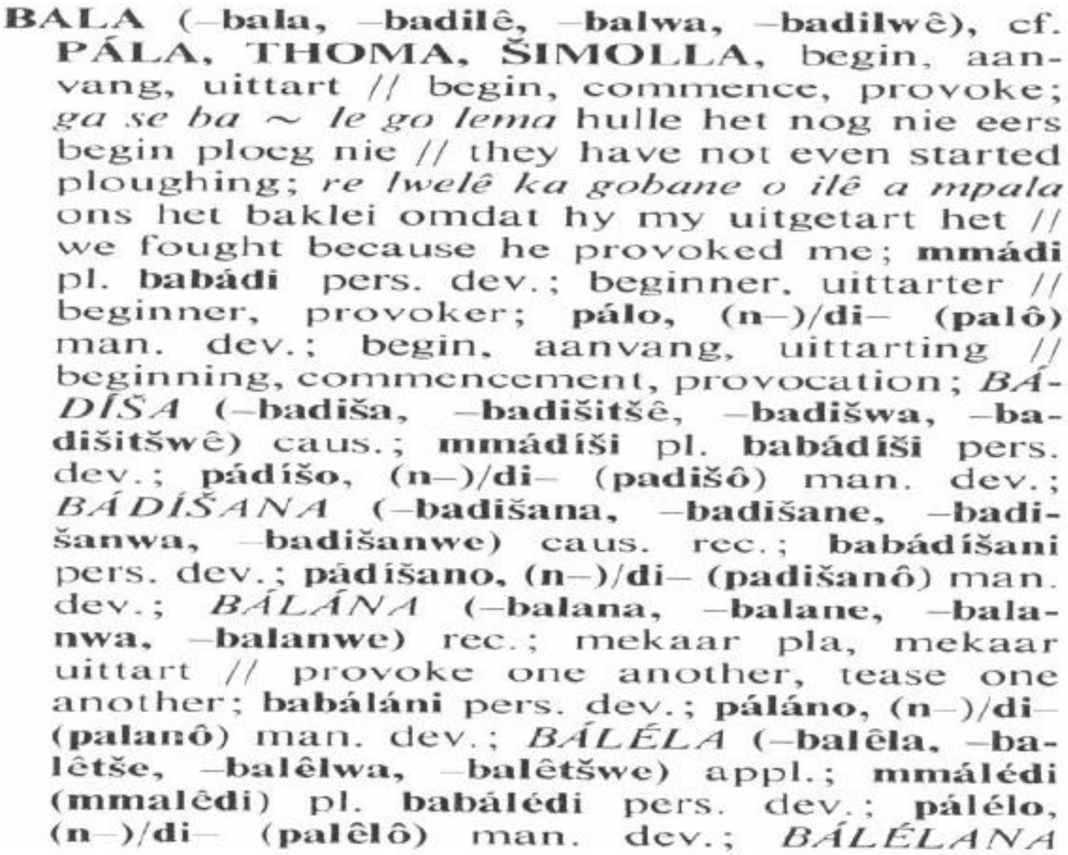
(kú)-hambelâna *v.* 1. go with each other, travel together. 2. be on good terms.

(kú)-hambísa *v.t.* 1. drive, cause or help to move, send off. 2. bid farewell to; send greetings to. 3. purge.

(kú)-hambisâna *v.* associate, accompany.

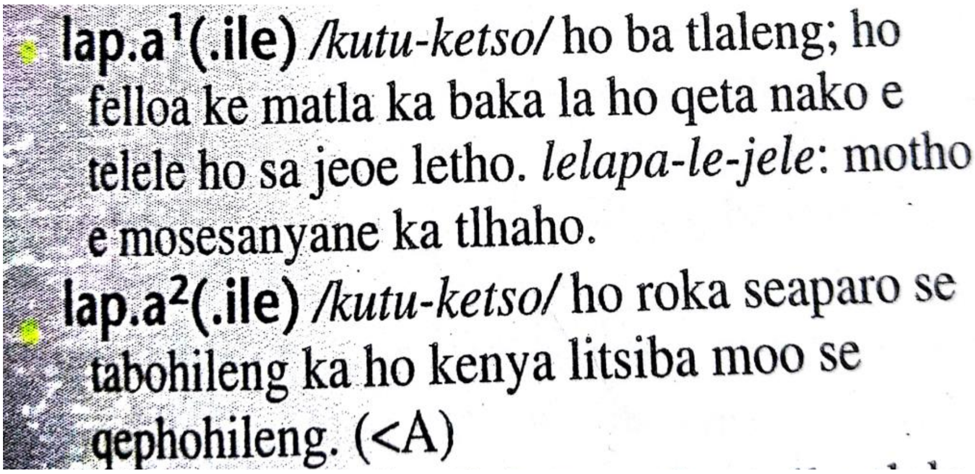
Figure 4.8.3: Example of left-expanded lemmatisation by Prinsloo (2015:157)

Another aspect to consider is whether a lumping or splitting approach should be used.



BALA (**–bala, –badilê, –balwa, –badilwê**), cf. **PÁLA, THOMA, ŠIMOLLA,** begin, aanvang, uittart // begin, commence, provoke; *ga se ba ~ le go lema* hulle het nog nie eers begin ploeg nie // they have not even started ploughing; *re lwelê ka gobane o ilê a mpala* ons het baklei omdat hy my uitgetart het // we fought because he provoked me; **mmádi** pl. **babádi** pers. dev.; beginner, uittarter // beginner, provoker; **pálo, (n–)/di– (palô)** man. dev.; begin, aanvang, uittarting // beginning, commencement, provocation; *BÁDÍŠA* (**–badiša, –badišitšê, –badišwa, –badišitšwê**) caus.; **mmádíši** pl. **babádíši** pers. dev.; **pádíšo, (n–)/di–** (**padišô**) man. dev.; *BÁDÍŠANA* (**–badišana, –badišane, –badišanwa, –badišanwe**) caus. rec.; **babádíšani** pers. dev.; **pádíšano, (n–)/di– (padišanô)** man. dev.; *BÁLÁNA* (**–balana, –balane, –balanwa, –balanwe**) rec.; mekaar pla, mekaar uittart // provoke one another, tease one another; **babáláni** pers. dev.; **páláno, (n–)/di– (palanô)** man. dev.; *BÁLÉLA* (**–balêla, –balêtše, –balêlwa, –balêtšwe**) appl.; **mmálédi (mmalêdi)** pl. **babálédi** pers. dev.; **pálélo, (n–)/di– (palêlô)** man. dev.; *BÁLÉLANA*

Figure 4.8.4: Example of lumping from Ziervogel and Mokgokong.



lap.a^{1}(.ile) */kutu-ketso/* ho ba tlaleng; ho felloa ke matla ka baka la ho qeta nako e telele ho sa jeoe letho. *lelapa-le-jele*: motho e mosesanyane ka tlhaho.

lap.a^{2}(.ile) */kutu-ketso/* ho roka seaparo se tabohileng ka ho kenya litsiba moo se qephohileng. (<A)

Figure 4.8.5: Example of splitting from Sethantšo sa Sesotho.

In figure 4.8.5, the lexicographer has differentiated between two meanings of the verb *lapa* 'to be hungry' or to 'patch clothes'. It is this separation or combination of meaning that separates the lumpers from the splitters in lexicography. Whereas, in figure 4.8.4, all meaning is combined in one article.

For the Sesotho language, the user needs to know the alphabetic system in figure 4.8.2 above. Depending on the theory used by the lexicographer, they might also need to know both ordinary alphabetic ordering and phonetic ordering for them to be able to find the words they are searching for in the dictionary, especially when the dictionary being compiled follows a phonetic ordering. However, not every person is used to or uses phonetic ordering (cf. figure 4.8.2) and phonetic rules; as a result, users are most likely to be left in dismay and confusion if they are to consult phonetic dictionaries. Consider the following quotes:

> …[a] phonetic dictionary orders expressions according to their pronunciation…, the order within the word list is not based on the written alphabet, due to the opinion that "spoken language, opposite to written language, is an independent domain whose structure should be analysed on its own principles…that way, an ordering is established which is closer related to phonetics and phonology, but this order is not a widely known one and untrained users will face problems locating the expression they are looking for. (Gouws, et al. 2013:798)

> The words in a dictionary are organized into several Index groups. The words within an Index group are ordered phonetically. The Phonetic Ordering Scheme is basically similar to the alphabetic ordering scheme except that symbols belonging to the same Phonetic Equivalence Group have the same phonetic order. Bandyopadhyay (1999:19)

4.9 Sesotho noun classes

Nouns are categorized into different noun classes which have different prefixes, as shown in table 4.9, where the class, prefix, subject concord, example and translation of example have been recorded.

Class	**Prefix**	**Subject concord**	**Example**	**Translation of example**
1	mo-	o	**mo**sadi	woman
2	ba-	ba	**ba**sadi	women
1a	-	o	ntate	father
2a	bo-	ba	**bo**ntate	fathers
3	mo-	o	**mo**se	dress
4	me-	e	**me**se	dresses
5	le-	le	**le**leme	tongue
6	ma-	a	**ma**leme	tongues
7	se-	se	**se**fate	tree
8	di-	di	**di**fate	trees
9	-	e	ntja	dog
10	di-	di	**di**ntja	dogs
Classes 11, 12 and 13 are mainly used in languages such as isiZulu and isiXhosa				
14	bo-	bo	**bo**hobe	bread
15	ho-	ho	**ho** tsamaja	to walk
16	-	ho	fatshe	down
17	ho-	ho	**ho**dimo	Up
18	mo-	o	**mo**se	abroad

Table 4.9: Sesotho Noun Classes (http://www.sesotho.web.za/nouns.htm)

Sesotho noun classes are divided into singular and plural forms. Lemmatising both singular and plural forms of nouns proved to be user-friendly and does not require any previous knowledge of the language on the side of the user. The most important thing is for them to know the alphabet of their language. (Prinsloo and De Schryver, 1999: 267)

> The Bantu languages are characterized by a nominal class system according to which nouns are sub-classified into different noun classes. These classes have a complex concordial and pronominal system, and complex word formation strategies by means of numerous affixes to verbal and nominal stems. (Prinsloo, 2012: 127- 128)

Van Wyk (1995) rejects the assumption that the stem tradition should be viewed as the more scientific method. That also makes sense, and this paper notes that, for Sesotho, it is empirical to use the word-based method because it is user-friendly and does not require that the users have an initial understanding of the morphology of words. With an example being De Schryver's Pukuntšu ya Sekolo (School Dictionary) (2007), a Northern Sotho-English Dictionary which used the word approach to lemmatise words. Full words and combinations are described, not part of words.

This notion makes even more sense because African languages, especially Sesotho, have a lot of bi-graphs, where two sounds are combined to make one sound (e.g. ***hl**aba*, 'stab' ***tj**o**tj**o* 'bribe' as well as tri-graphs where three sounds are combined to make one sound (e.g. ***tjh**elete* 'money', ***tsh**aba* 'afraid'), and it is important to take note of them. This way the user will know exactly where to look for words they want. This means that if a user wants to find the meaning of the word *mosebetsi* 'work' – he/she will have to look under headwords starting with "*mo*-" as the class prefix. And if they are looking for *tshaba* 'afraid', they will have to look under words starting with the tri-graphs '*tsh*-'. This alphabetic ordering is not haphazard and allows for the correct order of the Sesotho alphabet.

4.10 Macrostructural inconsistencies

There are certain macrostructural inconsistencies that have been picked up by different scholars, and some carry more weight than others. However, the ones that stood out are summarised by De Schryver and Prinsloo (2001: 376) in the following quote.

> 1. Inconsistencies on the relative length of alphabetical stretches, by treating certain sections of the lemma-sign list more exhaustively than others;

2. Inconsistencies regarding the creation of the lemma-sign list such as:

2.1. the omission of words most likely to be looked for, while words less likely to be looked for are included,

2.2. the partial treatment of lexical items belonging to a closed set,

2.3. the unequal treatment of various prefixes,

2.4. the absence of a policy to deal with productive versus non-productive suffixes,

2.5. the blind running of each stem through all possible verbal and nominal derivations, simply concatenating affixes, which results in serious doubts among mother-tongue speakers whether many of these derivations do exist,

2.6. the ad hoc handling of transparent versus non-transparent derivations;

3. Inconsistencies in terms of the choice of canonical forms.

All these inconsistencies were picked up in dictionaries which did not utilize a corpus and as a result De Schryver and Prinsloo (2000) concluded that African-languages lexicography needs urgently to be improved and has to use new and modern ways and tools in lexicography if they are to take their rightful place in the new millennium.

> African-language lexicographers, in other words, have no time left to rediscover the wheel. The challenge is thus to compile dictionaries for African languages according to the latest trends and most modern approaches in lexicography.
> (De Schryver and Prinsloo 2000:291)

These days, there are many ways that a lexicographer can simplify their work, with the user in mind. As such, the modern lexicographer is constantly looking for ways in which the dictionary can be improved to increase the success of information retrieval by target users. For the compiler of dictionaries for the African languages, the challenge is even greater since the lexicographer is the mediator between a complicated grammatical system on the one hand and the dictionary user on the other hand. (Prinsloo and De Schryver 2000:188) This essentially means that in every lexicographic process, African-languages lexicographers need to take a stand and start using modern ways of dictionary making that are friendly to the user while containing all the necessary information. Hence the user's needs have to take top priority in all lexicographic activities.

4.11 Conclusion

On the level of the macrostructure, corpora play a major role in lemma selection and the compilation of lemma lists for new dictionaries or the revision of lemma lists for existing ones. The major consideration is the frequency of occurrence of words in spoken and written texts, and this has a direct implication on inclusion or omission from the dictionary. Frequency lists supplement the intuition of the lexicographer and ensure that frequently used words most likely to be looked for by the target users are not accidentally omitted simply because they, in the view of Snyman et al. (1990: preface) did not cross the compilers way. In addition, frequency counts alert the lexicographer to words in the lemma list of an existing dictionary which are unlikely to be looked for, i.e., all words in the dictionary with zero occurrence in the corpus should be considered as candidates for omission.

This study has discussed the types of macrostructures found in dictionaries, namely the simple macrostructure and the complex macrostructure. Macrostructural components play a major role in dictionaries as they form the core of a dictionary. They include lemmas treated in the dictionary, the front matter and back matter, use of frequency tests, etc. Sesotho noun classes are divided into singular and plural forms and are complex. Lemmatising both singular and plural forms of nouns proved to be user-friendly as this does not require that the user has any previous knowledge of the language.

A lexicographer has an important role to play in dictionary making, which is to negotiate a complex interplay and overlap between (a) lemmatisation approaches (b) lemmatisation strategies, (c) lexicographical traditions, (d) nominal and verbal structures and (e) conjunctiveness versus disjunctiveness (Gouws and Prinsloo 2005:68). The ordering of lemmas is also important to note because it helps users to get the information they want. Even though there are dictionaries compiled using a phonetic alphabetical order, as compared to alphabetic ordering, the former is not user-friendly and makes it hard for users to extract information, while the latter is better because people are already aware of the alphabetic rules.

To compile a macrostructure that is of a high standard, it is important to use the latest trends and approaches in lexicography.

Chapter 5: The Microstructure

5.1 Introduction

Different scholars have different ways of defining the microstructure of a dictionary as can be seen in the following quotations. A microstructure is defined by Hartmann and James (1998: xii) as:

> The format of the entry, how information about the headword is provided and presented, and the appropriateness of the discourse structure of the entry for the benefit of the anticipated user.

Prinsloo and Gouws (2005:138) provide another nuance to the definition:

> Although the microstructure of a dictionary can informally be described as the set of entries in a dictionary article accompanying the lemma and presented as the treatment of the lemma, the term microstructure demands a much more precise interpretation, and this interpretation should be influenced by a number of other features.

However, Nielsen (1994:219) points out a difficulty in the definition of the microstructure by stating that:

> … the microstructure of a dictionary is the structure of an entire article. However, this poses some difficulties in relation to present study of lexicographic structures.

The last two quotes show that there is more to the microstructure than meets the eye. Their focus on precise interpretation of the entry influenced by a number of factors, together with the difficulties posed by defining the microstructure as being a structure of an entire article, gives an indication that the microstructure is not just part of the structure of a dictionary. It is rather a complex structure which focuses on the entry of the dictionary and other aspects that form part of the entry.

The microstructure is an important part of a dictionary because it is where all teaching and understanding of words takes place. It forms the base of the dictionary because it is where users go first when they encounter words they do not know or understand. As a result, it should be treated as a prerequisite, and must have all the necessary and important information. The microstructure is a space where a good lexicographer would tread carefully to spare the user unnecessary frustration and confusion. It is also a space where the lexicographer has to use his/her role to make sure that important words that are most likely to be looked for are lemmatised and treated accordingly.

There are different forms of microstructures, and a lexicographer has to pick one which suits his/her purpose and be consistent with it throughout the dictionary. Prinsloo and Gouws (2005:138) explain this condition:

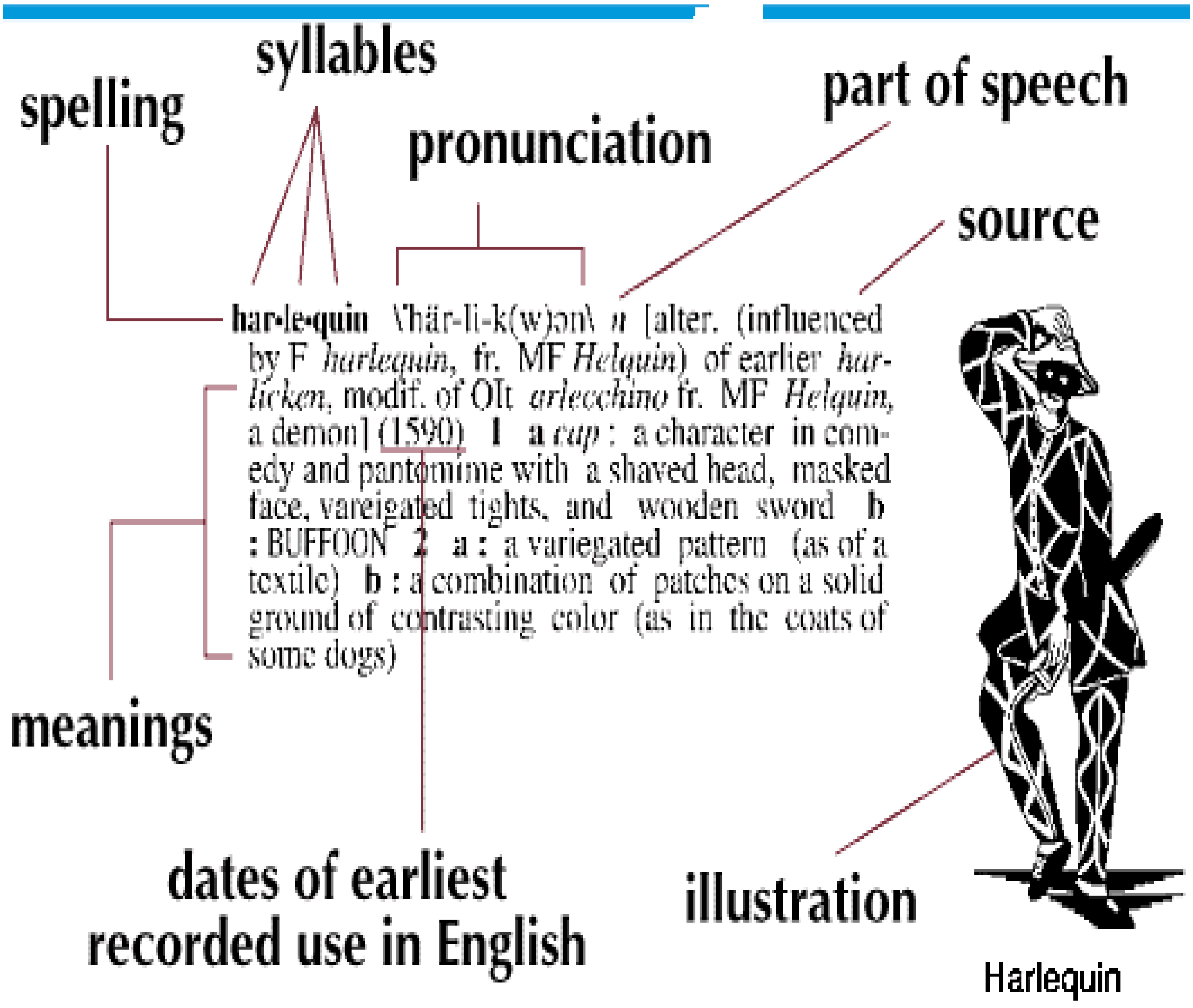
> The dictionary specific lexicographic process of each project must instruct the lexicographers with regard to the type of microstructure to be employed in the dictionary…the microstructure should be seen as an instrument to help achieve the genuine purpose of the dictionary. With regards to the types of microstructures, two major types can be distinguished, namely: the integrated and non-integrated microstructure which can further be divided into obligatory microstructure versus extended obligatory microstructures. The distinction between the two major types of microstructures, the integrated and non-integrated microstructure, is made on the grounds of the proximity and directness of the relation between each entry representing a paraphrase of meaning (in a monolingual dictionary) or each entry representing a translation equivalent (in a bilingual dictionary) and the supporting cotext entries representing illustrative examples in a specific article.

A dictionary entry is the first part that the user normally goes to when consulting a dictionary during their search for meanings or learning new words. The entry for a specific word is usually accessed through the lemma which represents the word and such lemmas are mostly presented in alphabetical order.

> When looking at the microstructure of a dictionary or when planning the microstructure of a dictionary, it is important that one should be well aware of the

> different types of entries to be included as microstructural elements. The term entry is used in different ways by different scholars, and one of the frequent uses of the term entry refers to the dictionary article. (Gouws and Prinsloo, 2005:115)

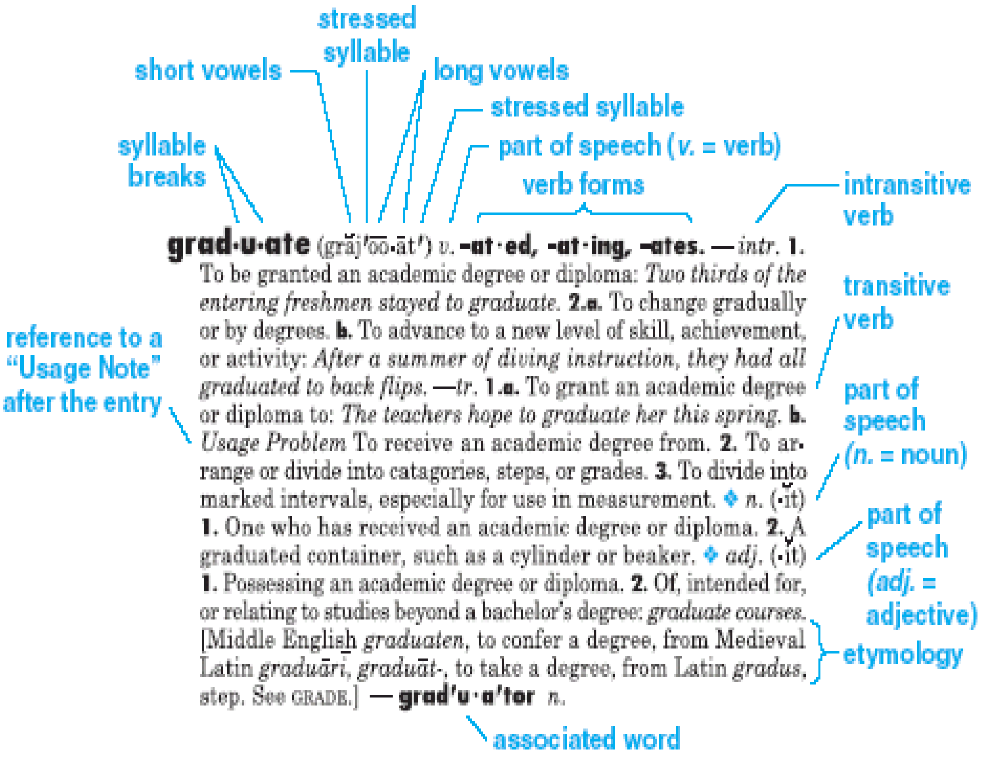
Different dictionaries focus on different aspects of the entry, but it is common for them to have a way of showing the user the spelling and syllable structure of the words. In figure 5.1.1 the article comprises of the following: spelling, syllables, pronunciation, part of speech, source, meanings, dates of earliest record in English and a picture illustration.



har·le·quin \ˈhär-li-k(w)ən\ *n* [alter. (influenced by F *harlequin*, fr. MF *Helquin*) of earlier *harlicken*, modif. of OIt *arlecchino* fr. MF *Helquin*, a demon] (1590) **1 a** *cap* : a character in comedy and pantomime with a shaved head, masked face, vareigated tights, and wooden sword **b** : BUFFOON **2 a** : a variegated pattern (as of a textile) **b** : a combination of patches on a solid ground of contrasting color (as in the coats of some dogs)

Figure 5.1.1: Illustration of the entry 'harlequin' (https://www.britannica.com/topic/dictionary/Kinds-of-dictionaries)

In most dictionaries the lemma is presented in **boldface**, definitions and translation equivalent paradigms in Roman numbers (i), and examples of usage in *italics*. These are, however, just typical conventions and can vary. Once the user finds the lemma, he/she can embark on a process of information retrieval by reading the different information types presented in the article of the lemma. It helps users understand the word by teaching them how it is pronounced, how it came about (etymology), how many syllables it has, its part of speech, illustrations, etc., which give a mental overview of what exactly is meant by the word together with many other aspects that different lexicographers decide to focus on.



grad·u·ate (grăj′o͞o·āt′) *v.* **-at·ed, -at·ing, -ates.** —*intr.* **1.** To be granted an academic degree or diploma: *Two thirds of the entering freshmen stayed to graduate.* **2.a.** To change gradually or by degrees. **b.** To advance to a new level of skill, achievement, or activity: *After a summer of diving instruction, they had all graduated to back flips.* —*tr.* **1.a.** To grant an academic degree or diploma to: *The teachers hope to graduate her this spring.* **b.** *Usage Problem* To receive an academic degree from. **2.** To arrange or divide into catagories, steps, or grades. **3.** To divide into marked intervals, especially for use in measurement. ◆ *n.* (·ĭt) **1.** One who has received an academic degree or diploma. **2.** A graduated container, such as a cylinder or beaker. ◆ *adj.* (·ĭt) **1.** Possessing an academic degree or diploma. **2.** Of, intended for, or relating to studies beyond a bachelor's degree: *graduate courses.* [Middle English *graduaten*, to confer a degree, from Medieval Latin *graduāri, graduāt-*, to take a degree, from Latin *gradus*, step. See GRADE.] — **grad′u·a′tor** *n.*

Figure 5.1.2: Illustration of the entry 'graduate' from American Heritage Dictionary of the English Language 4th edition cited in http://college.cengage.com

Figure 5.1.2 illustrates how different dictionaries focus on different aspects of the entry. It has different constituents (compared to figure 5.1.1), namely the spelling of the main entry as the point of departure, even though it is not stressed, the syllable breaks, types of vowels (short and long vowels), pronunciation where the stressed syllable is indicated, the part of speech with different verb forms (intransitive and transitive verb). The secondary meaning is also shown with its parts of speech together with its etymology and associated word. In the American Heritage Dictionary entry (figure 5.1.2), the dates of earliest recorded use in English and illustrations have not been supplied. The lexicographer of each dictionary did what they saw fit for their intended users.

Depending on the type of dictionary, each article will vary, in order to serve the needs of a particular user, and this point is reiterated by Fuertes-Olivera (2010:29):

> Both terminographers and lexicographers can offer tools which could present two or more dictionary articles in an alphabetical or systematic way depending on the needs of the user and of course, such a choice is made possible in the user interface.

The diagram below (figure 5.1.3) shows how a lexicographer goes about conceptualizing a dictionary. A dictionary entry should be able to answer the questions posed in this diagram for it to be deemed user-friendly.

- What does the word mean?
- Where does it come from?
- Which part of speech does it fall under?
- How is the word pronounced?
- How is the word used? And so on.

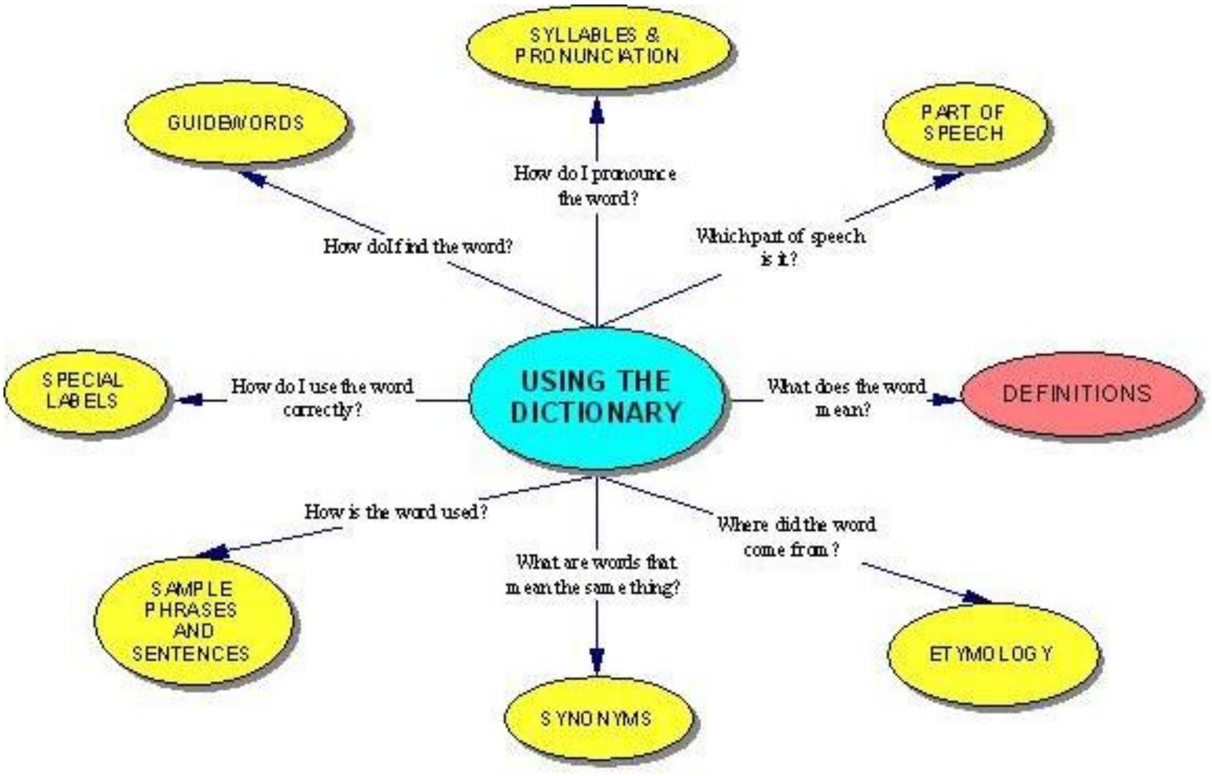

SYLLABLES & PRONUNCIATION
GUIDEWORDS
PART OF SPEECH
How do I pronounce the word?
How do I find the word?
Which part of speech is it?
SPECIAL LABELS
How do I use the word correctly?
USING THE DICTIONARY
What does the word mean?
DEFINITIONS
How is the word used?
Where did the word come from?
What are words that mean the same thing?
SAMPLE PHRASES AND SENTENCES
ETYMOLOGY
SYNONYMS

Figure 5.1.3: Dictionary Concept Map (http://highered.mheducation.com/sites/0073123587/student_view0/chapter4/ dictionaryconcept_map.html)

5.2 Corpora and the Microstructure

The microstructure gives information about the etymology meaning, semantic and syntactic information, noun class or part of speech, cross-references, pronunciation, examples of usage, etc., of words in a dictionary. Al-Jarf (1998:1) illustrates this notion in the following way:

> Modern dictionaries often include information about spelling, syllabication, pronunciation, etymology (word derivation), usage, synonyms, and grammar, and sometimes illustrations as well. This information is then structured in a way that makes it easily accessible and understandable, hence some online dictionaries have an audio function that enables the user to know and understand how a word is pronounced. The key here is 'users'. It is important to note that corpora help

> lexicographers achieve their goals and to write better dictionary articles. Meaning that corpora is an indispensable aid to dictionary compilation.

Hence Prinsloo and De Schryver (2000:310) suggest that “in order to respond to user needs, modern researchers must treat real language which is conveniently stored in electronic corpora”.

Electronic corpora gave the study of languages a new definition and purpose, and also gave rise to simpler ways of compiling and dealing with the microstructure. It has opened new possibilities for language scholars and language research. It has given scholars a chance to study words in context, and language in a more convenient and time-saving manner. As a result, it is important for African-language scholars to start reaping the benefits of the corpus and corpus studies and to take advantage of the possibilities they offer, especially in the compilation of microstructure.

In South Africa particularly, corpora have been built in the Department of African Languages of the University of Pretoria since the early 1990s, beginning with compiling a corpus for Sesotho sa Leboa, the *Pretoria Sesotho sa Leboa Corpus* (PSC), which grew from 156,000 running words or ‘tokens’ in 1990 to 8.7 million words currently and still growing. Thus, it is worth noting that by the end of the 1990s, corpora for other African languages were also built although they remained small compared to the current landmark in corpus creation for the African languages of South Africa, namely a 20.5 million-token corpus for isiZulu built at the University of KwaZulu-Natal, which is a significant milestone towards the intellectualization of isiZulu language. According to Khumalo (2017:18-19):

> The corpus has a Corpus Management System that has three critical suites that allows for wordlist and frequency searches, concordance function and keyword extraction. The application consists of two main parts, first is the server-side corpus query engine, which handles storage of the corpus and processes queries. Second is the client-side user interface, with which users interact.

Corpora allow for real language analysis and assist in writing better examples of usage, based on what is common and used by ordinary people. It has the ability to show real language use together with strings of words that normally occur together, together with their frequency of use, amongst other benefits. It is for this reason that the lexicographer has to use an electronic corpus instead of relying solely on intuition. Intuition is inadequate most of the time because it is human nature for people to forget, but corpus usage eliminates such occurrences. Prinsloo and Gouws (2005:34) states that "the chances of a dictionary compiler gathering all senses and sub-senses of highly polysemous words on the basis of intuition is questionable".

Utilization of a corpus requires the use of corpus query tools, and these two are inseparable because a lexicographer cannot extract that kind of information without a good corpus query tool. Corpus query tools manipulate the data in a corpus and provide the lexicographer with the much-needed results that allow him/her to draw important conclusions based on search queries. They also allow the lexicographer to present important aspects deemed necessary by the user. As a result, the correct use of a corpus and its corpus query tools is a winning formula for both the lexicographer and for the user. It is for this reason that Prinsloo and De Schryver (2000:311) said:

> For African Language lexicography to take its rightful place in the new millennium, [it] cannot stay on the side-lines when it comes to the active use of corpora to improve the quality of microstructural elements in the treatment of lemma signs.

5.3 Teacher-learner relationship

Compiling a microstructure with the use of a corpus enables the lexicographer to find ways of guiding the user on different aspects of the lemma, such as part of speech, etymology, examples usage, etc. It is safe to say one of the basic aims a lexicographer should achieve is to guide the user in respect of the properties / features/ characteristics/ uses /meanings of the lemma sign. (Prinsloo and De Schryver 2000:311). In this way, the lexicographer has to provide the user with enough information about the lemma and allow for a situation of learning on the part of the user and teaching on the part of the lexicographer, all through the dictionary. Laufer (1992:71) refers to the amount and nature of the required information as the urge '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 morphologically 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 ... The foreign language learner knows a much smaller number of words ... In many cases word knowledge is only partial, i.e. the learner may have mastered some of the word's properties but not the others.

Such an objective allows for a situation of learning on the part of the user, and teaching on the part of the lexicographer. All lexicographic prerequisites lie on the shoulders of lexicographers because they have to make sure that users are not left dismayed and confused, or misled by the dictionary, or feeling that they wasted their time in consulting it in the first place. It is therefore the sole responsibility of the lexicographer to ensure that the dictionary represents a language by lemmatising commonly used words, although it does not mean that rare words need to be disregarded. Nonetheless, it is important to note that "newly-emerging words, phrases and meanings need to be added in order to ensure that the dictionary remains current." (Rundell 2015:302).

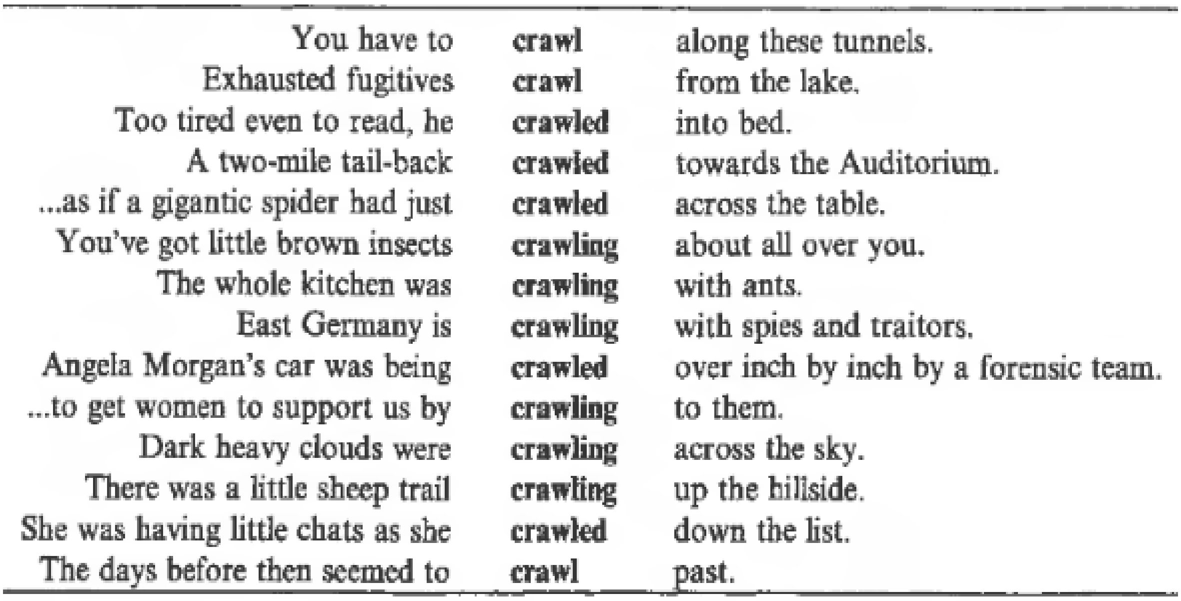
Prinsloo and De Schryver (2000:312) also agree that "… it is the aim of the lexicographer to present the microstructural components in such a way that enables the user to know the words at the entire information retrieval route". Knowing a word means knowing all its senses and sub-senses, its part of speech, pronunciation, etc., and what better way to attain all this than to use electronic corpora.

5.4 Sense distinctions and the corpus

A well written microstructure is one in which all senses of a word were considered, to account for all possible ways that a lemma can appear or be used. Corpora in this case, help the lexicographer to gather all sense distinctions of a word and assists them to deal with lemmas as they appear in real languages settings. Corpus query tools must be able to provide at least two basic outputs, namely, word frequency counts and concordance lines. "Corpus query outputs also help considerably in ascertaining whether all possible senses, or at least the main ones, of a particular lemma sign have been detected and treated.” Prinsloo and De Schryver (2000:312).

By sense distinctions, the lexicographer is merely looking at all possible ways a lemma can be used because he/she is always in doubt whether or not all relevant senses are covered in the definition or in the translation-equivalent paradigm. The lexicographers’ main aim is to lemmatise words that are most likely to be searched for, and all this can be achieved with the use of corpora and its corpus query tools. The correct use of a corpus and its corpus query tools has opened up a whole array of language study and offers an opportunity for the lexicographer to find all the necessary information during dictionary compilation, in a short space of time, from different genres of text, thereby helping him or her reach conclusions which could otherwise not have been reached.

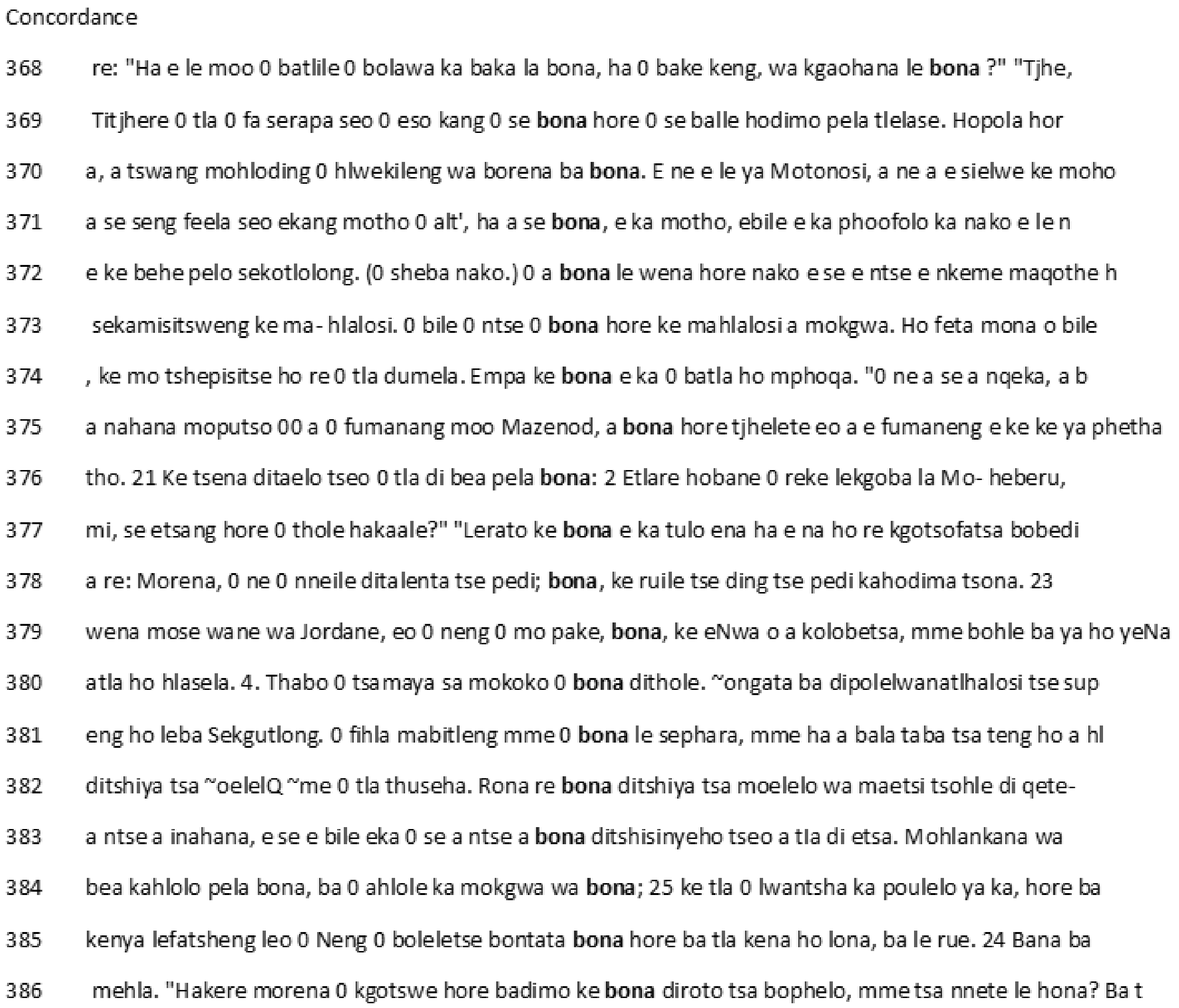
Consider the following example (figure 5.4.1) where different senses of the word *crawl* have been established using corpora. This way, the lexicographer can make conclusions speedily with just a few computer commands. A single glance at these lines is sufficient to detect the main senses of *crawl* such as ‘moving on hands and knees', 'time moving slowly', 'be overcrowded', etc. (Prinsloo and De Schryver 2000:312)



You have to	**crawl**	along these tunnels.
Exhausted fugitives	**crawl**	from the lake.
Too tired even to read, he	**crawled**	into bed.
A two-mile tail-back	**crawled**	towards the Auditorium.
...as if a gigantic spider had just	**crawled**	across the table.
You’ve got little brown insects	**crawling**	about all over you.
The whole kitchen was	**crawling**	with ants.
East Germany is	**crawling**	with spies and traitors.
Angela Morgan’s car was being	**crawled**	over inch by inch by a forensic team.
...to get women to support us by	**crawling**	to them.
Dark heavy clouds were	**crawling**	across the sky.
There was a little sheep trail	**crawling**	up the hillside.
She was having little chats as she	**crawled**	down the list.
The days before then seemed to	**crawl**	past.

Figure 5.4.1: Corpus lines for crawl* cited in Atkins, Rundell and Weiner (1997: slide 6abc2) and Prinsloo and De Schryver (2000:313)

Consider also a selection of the concordance lines for bona in Sesotho in figure 5.4.2.



Concordance

368 re: "Ha e le moo 0 batlile 0 bolawa ka baka la bona, ha 0 bake keng, wa kgaohana le **bona** ?" "Tjhe,

369 Titjhere 0 tla 0 fa serapa seo 0 eso kang 0 se **bona** hore 0 se balle hodimo pela tlelase. Hopola hor

370 a, a tswang mohloding 0 hlwekileng wa borena ba **bona.** E ne e le ya Motonosi, a ne a e sielwe ke moho

371 a se seng feela seo ekang motho 0 alt', ha a se **bona**, e ka motho, ebile e ka phoofolo ka nako e le n

372 e ke behe pelo sekotlolong. (0 sheba nako.) 0 a **bona** le wena hore nako e se e ntse e nkeme maqothe h

373 sekamisitsweng ke ma- hlalosi. 0 bile 0 ntse 0 **bona** hore ke mahlalosi a mokgwa. Ho feta mona o bile

374 , ke mo tshepisitse ho re 0 tla dumela. Empa ke **bona** e ka 0 batla ho mphoqa. "0 ne a se a nqeka, a b

375 a nahana moputso 00 a 0 fumanang moo Mazenod, a **bona** hore tjhelete eo a e fumaneng e ke ke ya phetha

376 tho. 21 Ke tsena ditaelo tseo 0 tla di bea pela **bona**: 2 Etlare hobane 0 reke lekgoba la Mo- heberu,

377 mi, se etsang hore 0 thole hakaale?" "Lerato ke **bona** e ka tulo ena ha e na ho re kgotsofatsa bobedi

378 a re: Morena, 0 ne 0 nneile ditalenta tse pedi; **bona**, ke ruile tse ding tse pedi kahodima tsona. 23

379 wena mose wane wa Jordane, eo 0 neng 0 mo pake, **bona**, ke eNwa o a kolobetsa, mme bohle ba ya ho yeNa

380 atla ho hlasela. 4. Thabo 0 tsamaya sa mokoko 0 **bona** dithole. ~ongata ba dipolelwanatlhalosi tse sup

381 eng ho leba Sekgutlong. 0 fihla mabitleng mme 0 **bona** le sephara, mme ha a bala taba tsa teng ho a hl

382 ditshiya tsa ~oelelQ ~me 0 tla thuseha. Rona re **bona** ditshiya tsa moelelo wa maetsi tsohle di qete-

383 a ntse a inahana, e se e bile eka 0 se a ntse a **bona** ditshisinyeho tseo a tla di etsa. Mohlankana wa

384 bea kahlolo pela bona, ba 0 ahlole ka mokgwa wa **bona**; 25 ke tla 0 lwantsha ka poulelo ya ka, hore ba

385 kenya lefatsheng leo 0 Neng 0 boleletse bontata **bona** hore ba tla kena ho lona, ba le rue. 24 Bana ba

386 mehla. "Hakere morena 0 kgotswe hore badimo ke **bona** diroto tsa bophelo, mme tsa nnete le hona? Ba t

Figure 5.4.2: Concordance lines for *bona*

Different senses and homonyms of the word *bona* emerge in the illustration above. This way, the lexicographer is able to provide plenty of examples of use of the user. For example, in concordance line 374, *ke mo tshepisitse ho re 0 tla dumela. Empa ke **bona** e ka 0 batla ho mphoqa. "0 ne a se a nqeka, a b.* the word ***bona*** is used as ‘see’, whereas in line 384 *bea kahlolo pela **bona**, ba o ahlole ka mokgwa wa **bona**; 25 ke tla 0 lwantsha ka poulelo ya ka, hore ha ba* means ‘before them’ and ‘their way’, respectively. On line 374, the ‘o’ in *bona* has a high tone whereas the ‘o’ of *bona* in line 384 has a lower tone. Tone marks the difference in these examples and therefore gives rise to homographs of the word *bona*, where *bona* in line 374 and in line 384 look alike but they are pronounced differently. The homonymity of *bona* is also shown in line 374 and line 377 respectively, where the word *bona* is spelled and pronounced the same. For example, in the sentence *ke mo tshepisitse ho re 0 tla dumela. Empa ke* ***bona*** *e ka 0 batla ho mphoqa. "0 ne a se a nqeka, a b* on line 374 and *mi, se etsang hore 0 thole haakale" "Lerato* ke ***bona*** *e ka tulo ena ha e na ho re kgotsofatsa bobedi.*

5.5 Corpora and definitions

Since the corpus era dawned upon the study of languages, lexicographers have been given the opportunity to write better dictionary definitions and articles. This means that users get exactly what they are looking for in the dictionary, and a happy user is a knowledgeable user. They are able to trust a dictionary as a source of information and are also able to consult it without any anxiety.

When it comes to definitions, it is important to have clear, understandable and well-structured information which will not leave the user with more questions than answers. It is therefore the responsibility of a lexicographer to deal with all aspects of the definition and translation-equivalent paradigm when compiling a dictionary and to focus his/her energy there because that is where the strength and core of most general dictionaries lie. According to Prinsloo and De Schryver (2000:315): "Corpus lines such as those presented in figure 5.4.2 are an excellent starting point for writing definitions and setting up translation equivalent paradigms".

Such definitions provide proper understanding of lemmas, together with properly written articles. Hence Cowie (2002:178) claims that:

> As a general rule, dictionaries are turned to as sources of information about meaning and to a lesser extent about spelling. This has long been the case, and as far as English dictionaries are concerned, is a generalisation that applies as much to foreign native users.

5.6 Corpora and collocations

Collocations are groups of words that normally occur together in one or more settings. They are regular groupings of words that help the lexicographer understand the different contexts in which word pairs occur. See the following definitions of collocations. According to the Oxford online dictionary (n.d.), a collocation is defined in the following manner:

> "The habitual juxtaposition of a particular word with another word or words with a frequency greater than chance. 'The words have a similar range of collocation'. A pair or group of words that are habitually juxtaposed. ''Strong tea' and 'heavy drinker' are typical English collocations'.

Nesselhauf (2005:11) points to the uses of collocation as follows:

> The term 'collocation' is used in widely different and often rather vague senses in linguistics and language teaching. The only common denominator is that the term is (at least mostly) used to refer to some kind of syntagmatic relation of words.

Other contributions to the definition of this concept include the following:

> Collocations might be described as words that are placed or found in a predictable pattern. Examples range from two-word combinations such as 'problem child' to extended combinations such as 'he is recovering from a major operation. These language patterns comprise much of speech and writing (Lewis, 2000).

The definitions above seem to have one thing in common, which is the idea that collocates are words often found close to each other. For example, in Sesotho, the words *ipona molato* 'to see one's fault' would be perfect. It would also not come as a surprise if these collocates were to have a high frequency rating because it is a predictable combination most likely to be understood by any Sesotho speaker.

However, Cowie (1998:191) gives a different opinion about collocations and contends that:

> There does not seem to be any clear cut, non-controversial definition of the term ‘collocation’. A vague definition would be to say that collocations are groups of words which frequently occur in combination with each other, however, this definition is totally unsatisfactory since it says nothing about the number of elements involved, the degree of frequency of occurrence or the classes of words which combine.

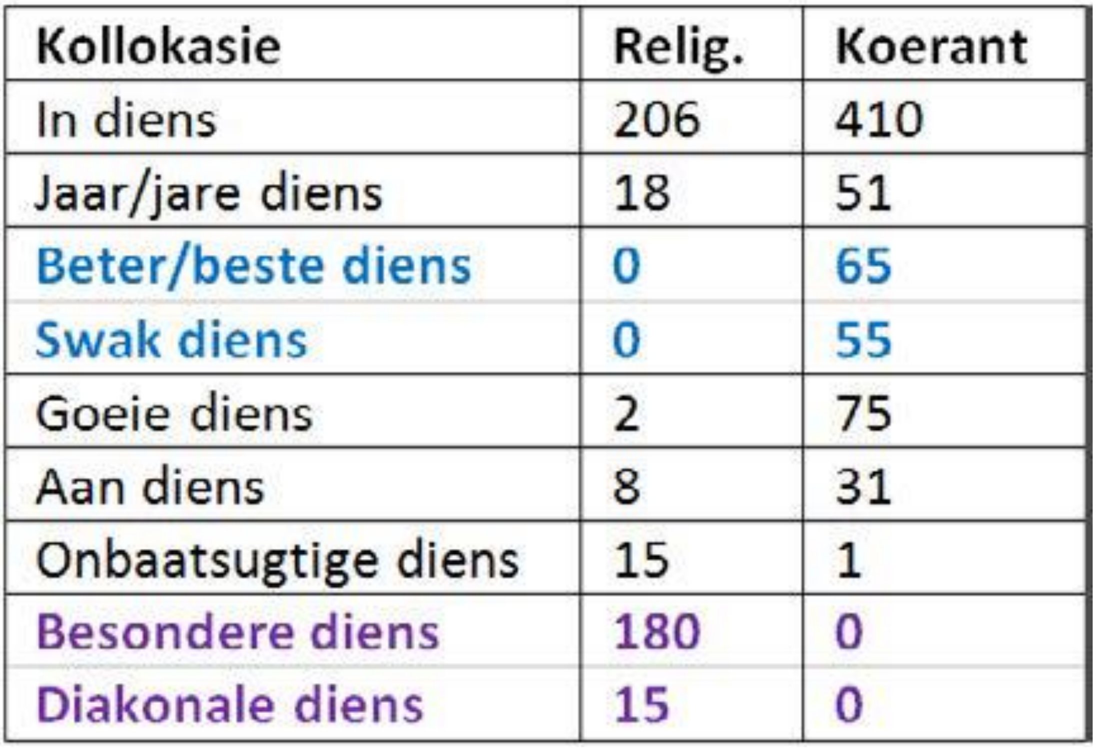
Proper lexicographic treatment of collocations enhances the quality of the microstructure and, together with frequency lists, they enable the lexicographer to understand the number of times certain words occur (together) and the order in which they occur. A single word does not have the same meaning as word pairs. For example, *ipona*, literally translated, is ‘see oneself’, but the moment *molato* (‘fault’) is added, the meaning of *ipona* totally changes. Now the meaning is no longer focused on *seeing oneself* but another aspect has been added, specifically ‘*see one's fault'*.

Bartsch (2004:11) emphasizes the point made earlier that mother tongue speakers will automatically understand predictable combinations in their languages:

> Proficient native language users are intuitively aware that some words in their language in some unspecified way tend to co-occur in relatively fixed and recurrent combinations...

Words that tend to occur together have the same meaning to speakers of the language, and when one word appears in a text or speech, its twin follows automatically. Having collocations is critical to lexicography because they offer known examples of usage to be used in the dictionary. Hence, Nugues (2006:87) notes that “in lexicography, extracting recurrent pairs of words – collocations – is critical to finding the possible contexts of a word and citing real examples of use.”

In figure 5.6.1, guidance in terms of frequency of use in religious versus newspaper corpora is given for collocations of the Afrikaans word *diens* (‘service’). The collocations in the diagram indicate exactly the number of times the word *diens*, appears in newspapers and religious books. Very often *in* will go together with *diens*, hence the high number of times it appears in both genres.



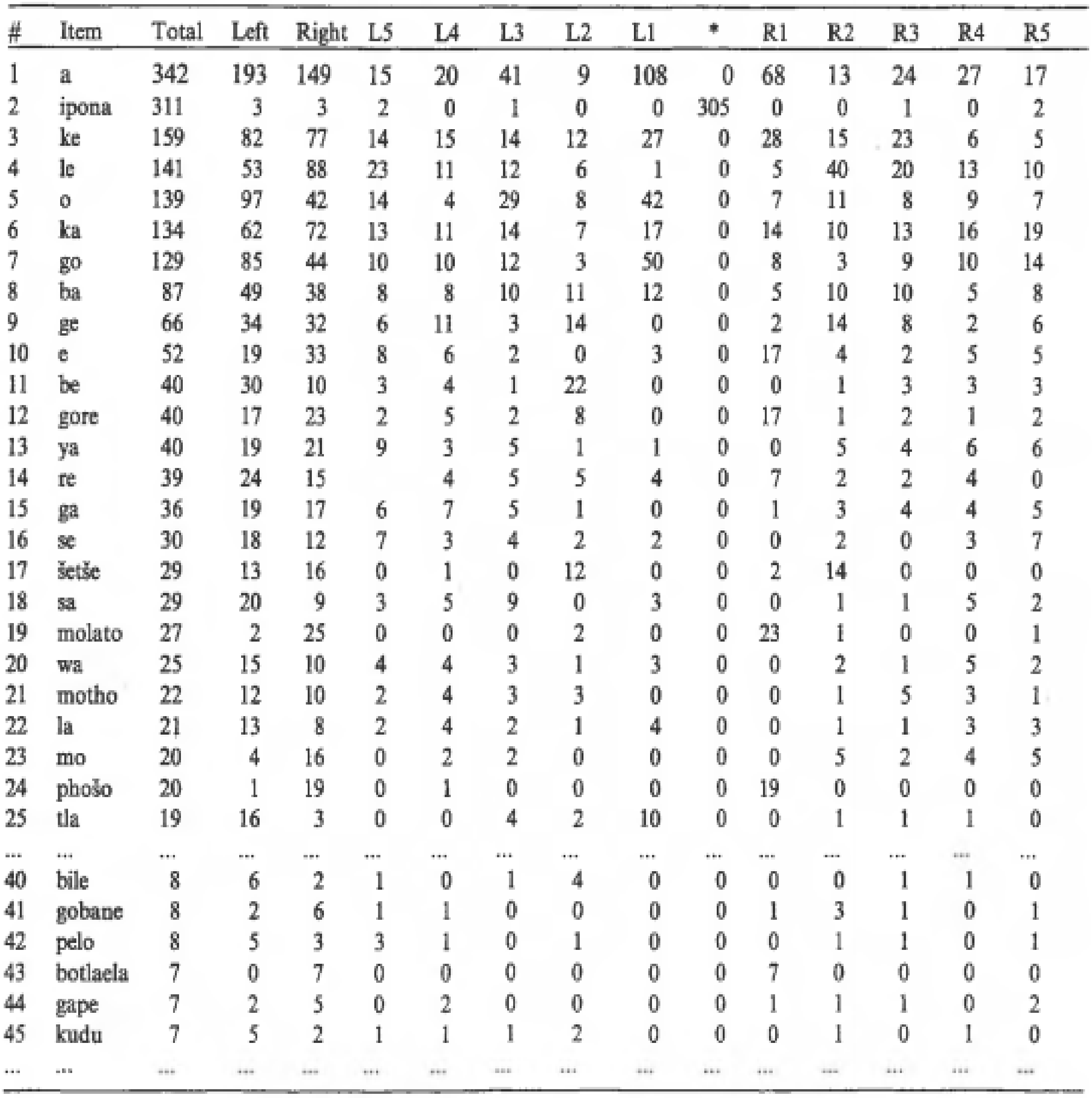
Kollokasie	Relig.	Koerant
In diens	206	410
Jaar/jare diens	18	51
Beter/beste diens	0	65
Swak diens	0	55
Goeie diens	2	75
Aan diens	8	31
Onbaatsugtige diens	15	1
Besondere diens	180	0
Diakonale diens	15	0

Figure 5.6.1: Domain-specific information (Prinsloo, Bothma and Heid 2013)

This is a painstaking task that that cannot under any circumstance be archived manually or through intuition. Once again, corpus query outputs provide valuable information to the lexicographer by indicating and computing statistics (as in figure 5.6.2) for collocates of *ipona* where the query tool was instructed to calculate its collocates. It is for this reason that Prinsloo and De Schryver (2000:315) state that:

> …if one instructs the corpus query tool to calculate and list collocations of the base term *ipona* 'see oneself' in order of their frequency, looking up to five places to the left and up to five places to the right, one sees that a grammatical word such as *molato* 'fault' collocated 27 times with *ipona* in the two horizon L5-R5. 2 collocates occur to the left of *ipona* and 25 to the right.

It therefore goes without saying that no mother tongue speaker could come close to presentations of collocations like the ones on figures 5.6.1 and 5.6.2 respectively on this form if they based their selection purely on intuition.

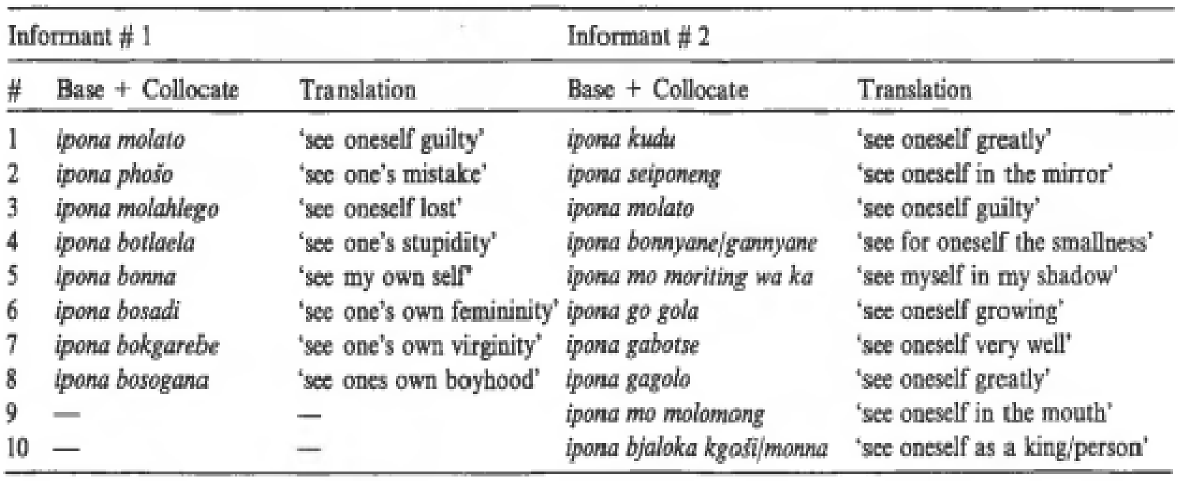


#	Item	Total	Left	Right	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	a	342	193	149	15	20	41	9	108	0	68	13	24	27	17
2	ipona	311	3	3	2	0	1	0	0	305	0	0	1	0	2
3	ke	159	82	77	14	15	14	12	27	0	28	15	23	6	5
4	le	141	53	88	23	11	12	6	1	0	5	40	20	13	10
5	o	139	97	42	14	4	29	8	42	0	7	11	8	9	7
6	ka	134	62	72	13	11	14	7	17	0	14	10	13	16	19
7	go	129	85	44	10	10	12	3	50	0	8	3	9	10	14
8	ba	87	49	38	8	8	10	11	12	0	5	10	10	5	8
9	ge	66	34	32	6	11	3	14	0	0	2	14	8	2	6
10	e	52	19	33	8	6	2	0	3	0	17	4	2	5	5
11	be	40	30	10	3	4	1	22	0	0	0	1	3	3	3
12	gore	40	17	23	2	5	2	8	0	0	17	1	2	1	2
13	ya	40	19	21	9	3	5	1	1	0	0	5	4	6	6
14	re	39	24	15		4	5	5	4	0	7	2	2	4	0
15	ga	36	19	17	6	7	5	1	0	0	1	3	4	4	5
16	se	30	18	12	7	3	4	2	2	0	0	2	0	3	7
17	šetše	29	13	16	0	1	0	12	0	0	2	14	0	0	0
18	sa	29	20	9	3	5	9	0	3	0	0	1	1	5	2
19	molato	27	2	25	0	0	0	2	0	0	23	1	0	0	1
20	wa	25	15	10	4	4	3	1	3	0	0	2	1	5	2
21	motho	22	12	10	2	4	3	3	0	0	0	1	5	3	1
22	la	21	13	8	2	4	2	1	4	0	0	1	1	3	3
23	mo	20	4	16	0	2	2	0	0	0	0	5	2	4	5
24	phošo	20	1	19	0	1	0	0	0	0	19	0	0	0	0
25	tla	19	16	3	0	0	4	2	10	0	0	1	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
40	bile	8	6	2	1	0	1	4	0	0	0	0	1	1	0
41	gobane	8	2	6	1	1	0	0	0	0	1	3	1	0	1
42	pelo	8	5	3	3	1	0	1	0	0	0	1	1	0	1
43	botlaela	7	0	7	0	0	0	0	0	0	7	0	0	0	0
44	gape	7	2	5	0	2	0	0	0	0	1	1	1	0	2
45	kudu	7	5	2	1	1	1	2	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 5.6.2: Collocates of the base *ipona* (with horizons L5-R5) in PSC

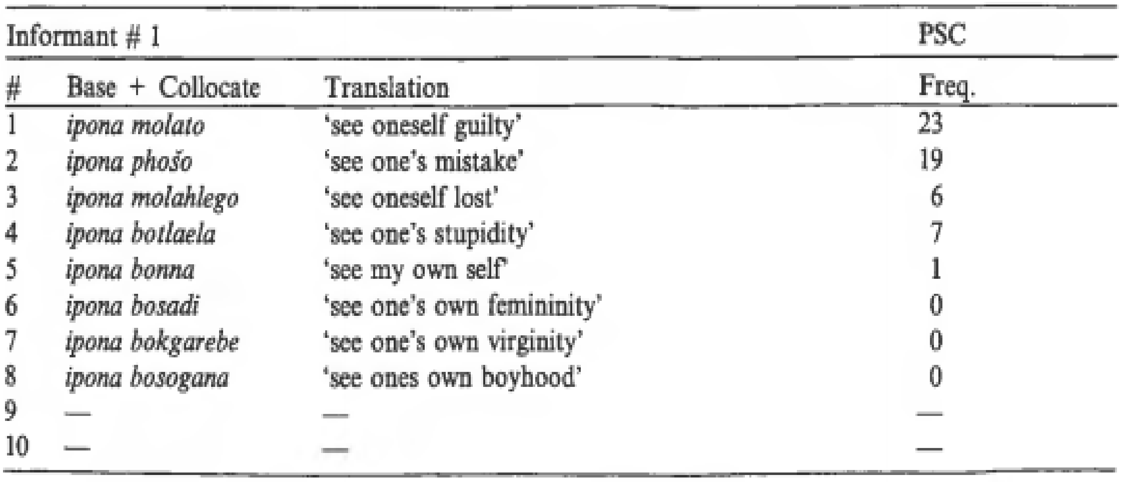
Such processed data of collocations allow for an in-depth study of word combinations, and play a major role in assisting the dictionary compiler to understand how words are used in context – even to the level of giving domain-specific guidance as described by Prinsloo, et al, above. Such information also provides for good examples which can be included in the article(s) of the collocator(s). They allow for comparisons of words to be done between mostly and widely used words or word strings, and also give an indication of what must be lemmatised. Also, because corpus studies are electronic, they have strong and in-depth data (especially when compared to speaker intuition) and are not affected by human error.

Refer to figures 5.6.3 and 5.6.4 as investigated by Prinsloo and De Schryver (2000:317). They reveal that intuition is not without errors and cannot be trusted. Different speakers will come up with different collocates and, while some sentiments could be shared, there are some instances where a break in the pattern can be seen. There exist differences in intuition on informant #1 and #2, which corpora easily solve.



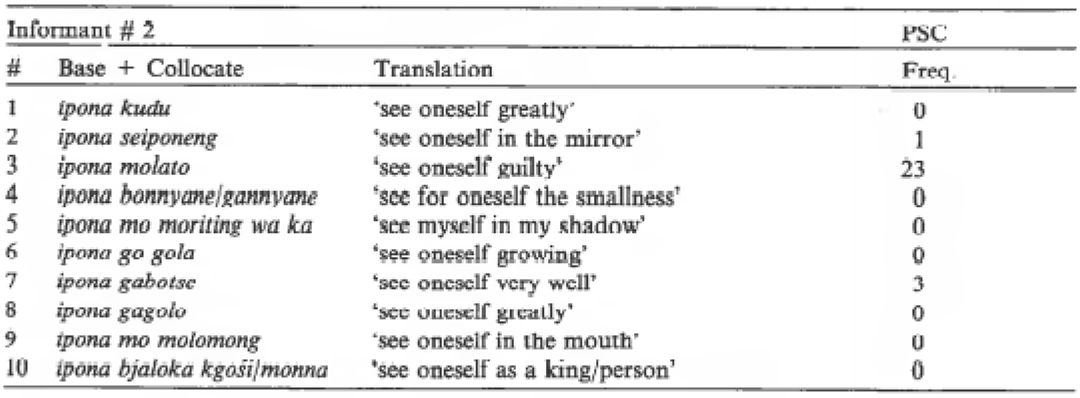
Informant # 1			Informant # 2	
#	Base + Collocate	Translation	Base + Collocate	Translation
1	*ipona molato*	'see oneself guilty'	*ipona kudu*	'see oneself greatly'
2	*ipona phošo*	'see one's mistake'	*ipona seiponeng*	'see oneself in the mirror'
3	*ipona molahlego*	'see oneself lost'	*ipona molato*	'see oneself guilty'
4	*ipona botlaela*	'see one's stupidity'	*ipona bonnyane/gannyane*	'see for oneself the smallness'
5	*ipona bonna*	'see my own self'	*ipona mo moriting wa ka*	'see myself in my shadow'
6	*ipona bosadi*	'see one's own femininity'	*ipona go gola*	'see oneself growing'
7	*ipona bokgarebe*	'see one's own virginity'	*ipona gabotse*	'see oneself very well'
8	*ipona bosogana*	'see ones own boyhood'	*ipona gagolo*	'see oneself greatly'
9	—	—	*ipona mo molomong*	'see oneself in the mouth'
10	—	—	*ipona bjaloka kgoši/monna*	'see oneself as a king/person'

Figure 5.6.3: 'Intuitive frequency' of the top ten collocates of the base ipona that collocate immediately to the right of *ipona* by Prinsloo and De Schryver (2000:317)



Informant # 1			PSC
#	Base + Collocate	Translation	Freq.
1	*ipona molato*	'see oneself guilty'	23
2	*ipona phošo*	'see one's mistake'	19
3	*ipona molahlego*	'see oneself lost'	6
4	*ipona botlaela*	'see one's stupidity'	7
5	*ipona bonna*	'see my own self'	1
6	*ipona bosadi*	'see one's own femininity'	0
7	*ipona bokgarebe*	'see one's own virginity'	0
8	*ipona bosogana*	'see ones own boyhood'	0
9	—	—	—
10	—	—	—

Figure 5.6.4: Informant #1's intuition compared to PSC by Prinsloo and De Schryver (2000:317)



Informant # 2			PSC
#	Base + Collocate	Translation	Freq.
1	*ipona kudu*	'see oneself greatly'	0
2	*ipona seiponeng*	'see oneself in the mirror'	1
3	*ipona molato*	'see oneself guilty'	23
4	*ipona bonnyane/gannyane*	'see for oneself the smallness'	0
5	*ipona mo moriting wa ka*	'see myself in my shadow'	0
6	*ipona go gola*	'see oneself growing'	0
7	*ipona gabotse*	'see oneself very well'	3
8	*ipona gagolo*	'see oneself greatly'	0
9	*ipona mo molomong*	'see oneself in the mouth'	0
10	*ipona bjaloka kgoši/monna*	'see oneself as a king/person'	0

Figure 5.6.5: Informant #2's intuition compared to PSC by Prinsloo and De Schryver (2000:317)

To back up the idea that intuition cannot be trusted, Prinsloo and De Schryver (2000:317) maintain that:

> Firstly, relying only on intuition, Informant # 1 listed four of the 10 most frequently occurring collocations namely *ipona molato, ipona phošo, ipona molahlego* and *ipona botlaela*. Even more significant is the fact that these four were the first that came to mind. It is, however, also apparent that intuition let the informant down from rank number 5 onwards, since the latter four suggested, namely *ipona bonna, ipona bosadi, ipona bokgarebe* and *ipona bosogana* did not, with the exception of

> *ipona bonna*, even occur once in the corpus… it is clear that intuition as to the typical collocates of the base *ipona* let Informant # 2 down. The first two collocates listed, as well as seven others further below, have almost zero occurrence in the corpus.

The use of corpora is also valuable because they allow for proper examination of different clusters of words together with their occurrences. Corpora give the dictionary compiler an indication of the frequency of use of words, together with information on words that normally go together. The word clusters are just a group of words and Scott (1996:35) refers to them as “words which are found repeatedly in each other’s company…” This tight relationship gives the lexicographer an indication on the extent of use of these words and how they should be dealt with in the dictionary-making process.

5.7 Corpora and the selection of examples

Corpora and corpus query tools can be used to find frequently occurring word pairs and examples used in natural settings. As language is fluid and changes over time, corpora are able to extract the most recent examples of use, thereby giving the user current examples in use. By combining different corpus tools, it is possible to find true examples of usage to be included in the dictionary. In this way, the user is exposed to an array of examples which are easy and comprehensible. This argument is therefore supported by Prinsloo and De Schryver (2000:326) when they say:

> …with good query tools at his/her disposal, the lexicographer can combine the output of different tools such as word-frequency counts and concordance line screens. For instance, by means of frequency counts, the lexicographer can determine that the most frequently used verb in Kiswahili is *kusema* ‘say; speak’ and that the most frequent inflection of that verb is *alisima* ‘he/she said; he/she spoke.

The combination of such corpus query tools shows that a lot can be achieved through the use of the benefits of corpora. Hence it is up to African-language lexicographers to harness and utilise corpora to allow for better understanding of languages, so that they are able to write better dictionaries and to find better ways to lemmatise words.

Prinsloo and De Schryver (2000:327) outline the benefits of the corpus data:

> Finally, with all the available corpus data, it is now very easy for the lexicographer to select a typical and natural example of usage for inclusion in the dictionary by simply glancing at the output of one or more concordance-line screens.

Corpus-based lexicography has opened up new ways of dealing with the macrostructure and has led to better research and understanding of meanings of articles, together with ways on how to approach the microstructural aspects from the user's perspective. The opportunity to write better and improved dictionaries through corpora is a total benefit for lexicographers and language scholars in general because important aspects – such as time constraints and getting all senses of a word – are not an issue.

5.8 Conclusion

Microstructure is defined in different ways by different scholars, but the common idea is that it focuses on the entry or what other scholars refer to as a dictionary article, consisting of the lemma and the treatment of the lemma. What should be included, and the kinds of questions each entry has to answer, are all dependent on the type of dictionary. Corpora help lexicographers to enhance the quality of treatment and to achieve their goals by ensuring that all microstructural elements are incorporated in a way that is convenient, time saving and user-friendly. Corpora draw on different genres and uses of words, as a result they help lexicographers find the different uses, etymologies, parts of speech, etc., of words which would otherwise be forgotten if lexicographers were to rely only on their instincts and intuition.

Prinsloo and De Schryver (2000:317) gave examples showing that intuition can be insufficient and that the use of corpora enhanced the quality of treatment in respect of collocations and examples of usage. Corpora and corpus query tools can be used to find frequently occurring word pairs and examples often used in natural settings, and the use of such tools also allows for better dictionary making. However, it is important to note that African scholars need to find ways to increasingly utilize this tool that has revolutionized language studies and research and made it possible for languages such as English to be studied and elevated.

Sesotho-language researchers and scholars need to use corpora to elevate their language and give it a chance to be used in socio-economic, political and educational spheres, and corpora will help Sesotho lexicographers to compile meaningful dictionaries. They have to leapfrog to the new way of doing things and also understand that it is possible for African languages generally to be equalled to other languages of the world, even though most of them have limited resources. The analysis of microstructures in Sesotho has inconsistencies which could have been avoided if corpora were used.

Chapter 6: Analysis of English Online Dictionaries

6.1 Introduction

The aim of this chapter is to critically analyse and evaluate two randomly selected e-dictionaries for English, namely the free Collins English-French online dictionary and the free Macmillan English Dictionary (MED). As language resources, they serve different purposes for different users, depending also on the skills users have. Collins has been publishing educational and informative books for almost 200 years and has been at the forefront of corpus studies, corpus-based lexicography and dictionary compilation, with the help of Sinclair who spearheaded corpus projects back in the day. Compare, for example, sources such as the Oxford Dictionary (both print and online), Cambridge Advanced Learner's Dictionary, Longman Dictionary of Contemporary English, and other big dictionaries. What is appealing about online dictionaries is that a user does not have to carry a big heavy book or to page through many pages to find the information they are looking for. The down side of e-dictionaries is their dependence on internet connection for accessibility purposes. Internet access has financial implications. The fact that space limitations are not an issue in online dictionaries is a massive improvement in dictionaries and has enabled editors and lexicographers to use their skills fruitfully without any words being omitted or some words given preference over others. This idea is echoed by Prinsloo (2005:12):

> These dictionaries can be utilized to their full capacity in terms of true electronic features… Whether online or on CD-ROM, such dictionaries present a new world of exciting electronic features.

6.2 The online dictionary

Online dictionaries have provided teachers, students and ordinary people with the ability to search for words even when mobile. All that users need to access online dictionaries is a computer or smartphone and internet connection. In the age of technology, where everything is now digital, where the use of smartphones and computers has increased, it becomes easy for users to manoeuvre through their cellphones or computers by clicking the *Google* search engine and searching for any word they have acquired or want to know more about. In other words, with just a click of a search button, the user is presented with a list of dictionaries or other references that contain the word. They easily go to the dictionary of their choice and, if they feel they are not getting what they want, they can easily go back and click on other dictionaries. The most praise-worthy features of online dictionaries are their ability to store an unlimited number of entries and be regularly updated. Lingling and Hai (2015:192-193) claim that:

> Due to the rapid development of the internet, smart phones and iPads, ways of obtaining information have been diversified. With the easy and widespread access to the internet, dictionary users can consult online dictionaries by hovering the mouse over a word on computer screens or clicking on their smartphones and iPads wherever they go.

As the shift from printed dictionaries to electronic dictionaries gained momentum, some big dictionary publishers like Macmillan discontinued paper-based dictionaries and concentrated only on their online dictionaries. They are using opportunities made available by the internet to reach their users and have been very successful. This idea is explored by the following quotations:

> Nearly all major traditional dictionaries now have online versions, whether partial or full, paid or free: they simply cannot afford to lose this battlefield and have tried to develop leading roles in providing language reference products and services on the Internet. (Lan 2005:16)

> Macmillan English Dictionary, one of the ‘Big Six’, has ceased the publication of paper dictionaries in 2013 and decided in favour of an online version only. Together with the multi-functions provided by online dictionaries, dictionary users can learn more about the latest revisions and new words editors can update entry information every few months. (Lingling and Hai 2015:193)

According to Pasfield-Neofitou (2009), there are different types of online dictionaries. There is the word dictionary, the glossary (including thesaurus), and the translators – and each of them has different uses for different users. In word dictionaries, users input a word and get a list of possible meanings (and other word-related information). In glossaries, users paste a text and get glosses for individual words/phrases; and in translators, users paste a text and get a translation of the whole text.

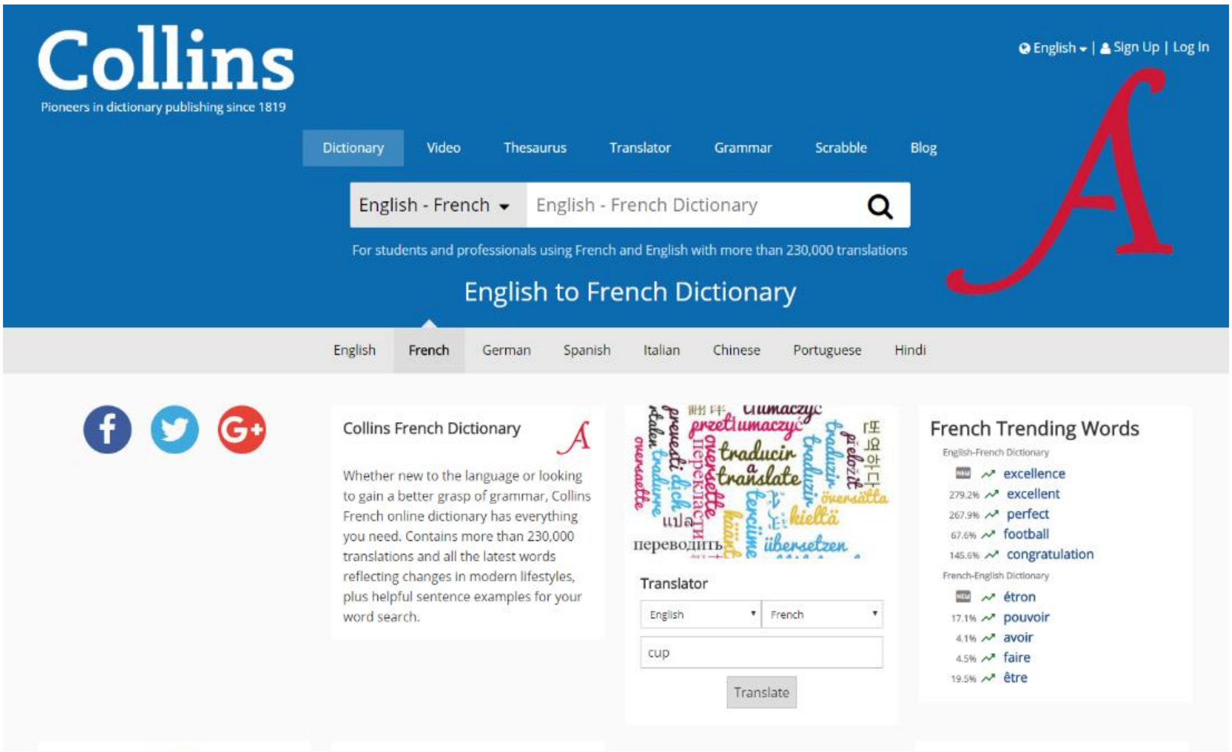
> Online dictionaries also differ in that some are paid, some partially paid while others are free to use. However, the paradox of online dictionaries is that those we pay for will have limited hits while those that are free, especially if good, will be jammed with callers. Lan (2005:16)

Online dictionaries have no space restrictions because they draw data from a diverse platform on the internet, and in some instances draw data from other dictionaries, allowing for multiple consultations. They are great time savers because they respond quickly to queries and their quality has improved over the years while speed in providing an answer has increased; it is possible to look up a word while working on one’s computer. There is also no doubt that new technology has speeded up searches and reduced waiting time, worry, and waste, and the birth of the broadband service provided by internet companies has enabled users to make the most of high-speed access to the internet. Lan (2005:18).

However, with all this said, the disadvantage of online dictionaries is the high cost of internet access. Without the internet, the user will not have access to any of the online dictionaries. In instances where internet coverage is poor, especially in rural and underdeveloped areas, access to such dictionaries is still a problem. Also, what happens when a user is busy searching for a word and their data gets depleted? This means that the user will be left frustrated and will not be able to reach their goal of dictionary use. In as much as online dictionaries have their pros and cons, they have revolutionised dictionaries in general and some publishers have even gone as far as discontinuing their paper-based dictionaries.

6.3 Analysis of the Collins English-French Online Dictionary

This dictionary is an English-to-French dictionary compiled for students and professionals with varying language levels. It states that it has more than 230,000 translations that help the user in understanding words they are looking for. In the screenshot of the Opening window on Collins (see figure 6.3.1) it can be noted that on the far left there are three icons which are *Facebook*, *Twitter* and *Google*, which shows that it is part of the broader digital community. There are different boxes promoting the learning of English-French and vice versa, while the last three situated at the bottom of the page are dictionaries in English-French. The first three are of utmost importance because they focus on the learning itself. (See figure 6.3.1). The first box introduces the Collins English Online Dictionary and is titled *Collins French Dictionary*. This introduction reveals the vision and aims of this dictionary. It gives reasons on why a learner of English-French should use this particular dictionary. This idea is expanded when the compilers emphasize: “whether new to the language or looking to gain a better grasp of grammar, Collins French Dictionary has everything you need”. (Collins n.d.)



For students and professionals using French and English with more than 230,000 translations

English to French Dictionary

Collins French Dictionary

Whether new to the language or looking to gain a better grasp of grammar, Collins French online dictionary has everything you need. Contains more than 230,000 translations and all the latest words reflecting changes in modern lifestyles, plus helpful sentence examples for your word search.

cup

French Trending Words

English-French Dictionary

NEW	excellence
279.2%	excellent
267.9%	perfect
67.6%	football
145.6%	congratulation

French-English Dictionary

NEW	étron
17.1%	pouvoir
4.1%	avoir
4.5%	faire
19.5%	être

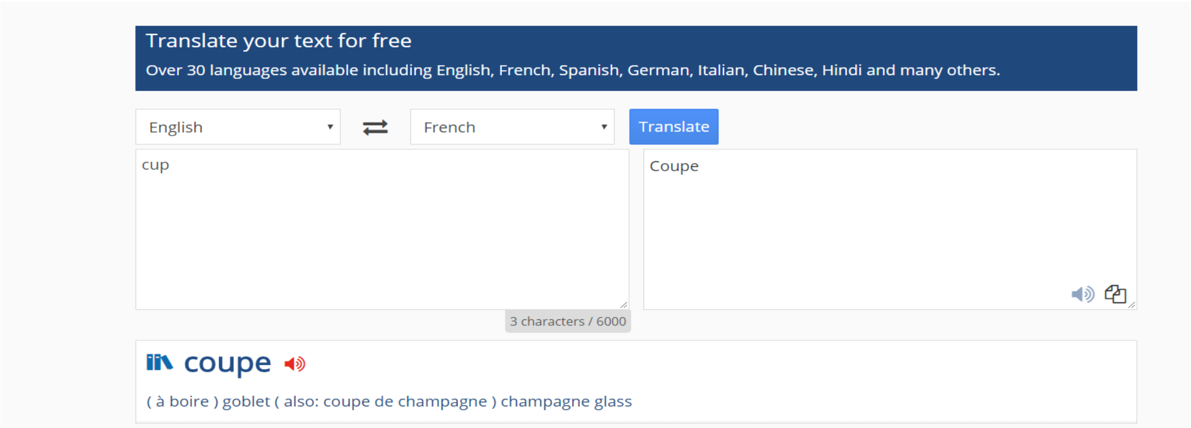
Figure 6.3.1: Opening window in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)

The dictionary emphasizes the notion that it will help the user to learn and understand the French language regardless of whether they are new to the language, or not. This dictionary contains more than 230,000 translated words; thus, it makes it easier for new learners/users to have a large repertoire of word references and to get a better understanding of them. This also means that usage of such words will also become easier with the right guidance. The inclusion of all the latest words in current use makes this particular dictionary a trustworthy source of information. The user can therefore rely on it. Another important aspect is the dictionary’s ability to present the user with helpful sentence examples relating to the searched words, which is vital in knowing and understanding a language.

The second box is for translation, and that is where all translation activities take place. It is titled *Translator* and reveals the teacher-learner relationship that exists between dictionaries and users, whether in dictionaries for special, general or any other purpose. A substantial part of the learning potential of dictionaries is to not only look up words in a specific theme but to also learn the translation equivalents, according to Prinsloo (2016:222). This box is one of the most important, as it is the one a user needs for a translation.

It has the word Translator in different colours to distinguish the different languages in which this Collins English Dictionary is available, and when you hover the cursor over that colourful box, the words ‘translate your text’ appear. This feature gives the user an idea on how to use the dictionary for translation purposes, and it also has other boxes with different functions which will be discussed below, with the first being the source language, which is English.

The moment the user types in a word (*cup* in this example), the following screen appears (see figure 6.3.2). The user is given the translation equivalent of the word *cup* which is *coupe* in French. There is also an icon of a speaker that, when clicked on, gives the pronunciation of the word *cup* in French, and this happens for all the words searched. There is a dropdown menu on both the ‘English’ and ‘French’ icons, and when the user clicks them, a variety of other languages appear, which could be source languages or target languages depending on the user.



Translate your text for free

Over 30 languages available including English, French, Spanish, German, Italian, Chinese, Hindi and many others.

English ⇄ French

cup

Coupe

coupe

(à boire) goblet (also: coupe de champagne) champagne glass

Figure 6.3.2: Translation of cup in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)

Another interesting fact about this dictionary is its ability to translate full phrases, like in the case of the phrase *I love you*, as illustrated in figure 6.3.3. Note that the pronunciation icon is also present here. The dictionary also gives the user other ways of using the subject pronoun ‘I’. When the words *I love you* are typed in the translation box the result *Je t'aime* appears on the box on the far right. This feature takes us back to the claim in their introduction – that whether a user is a novice to the language or is simply looking to gain a better grasp of grammar, Collins French online dictionary has everything they need. However, their choices of examples using *I'll* (which is the short form of *I will* or *I shall*), and *I'm* (which is the short form of *I am*), distinguishes between the different ways in which *I* is used. Consider the following:

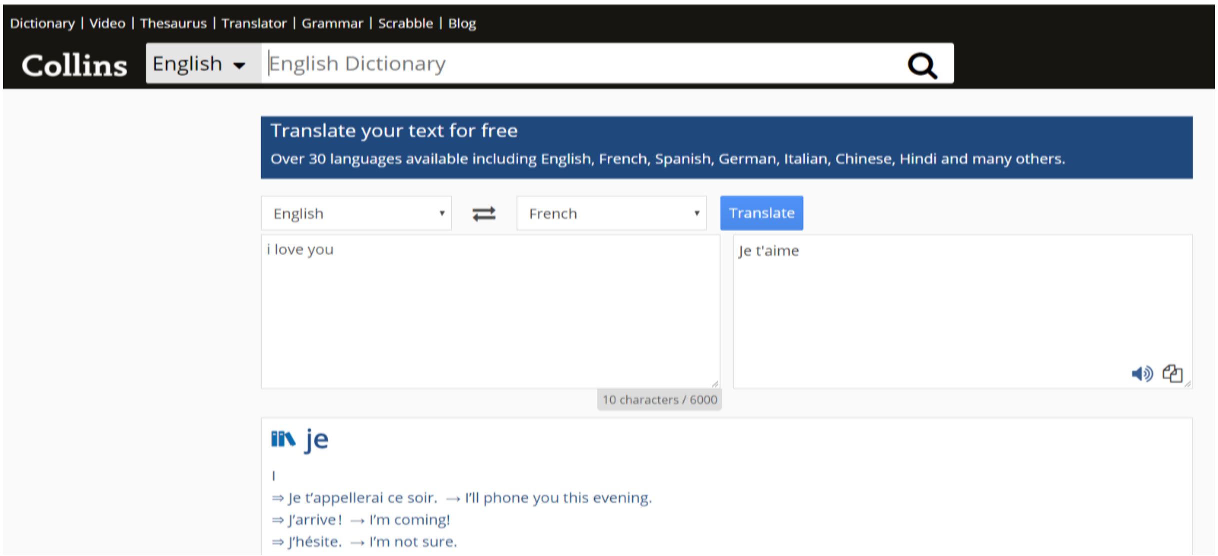
Je

I

⇒*Jet'appellerai ce soir*. →I'll phone you this evening.

⇒*J'arrive!* →I'm coming!

⇒*J'hésite.* →I'm not sure.


Dictionary | Video | Thesaurus | Translator | Grammar | Scrabble | Blog
Collins
English
English Dictionary
Translate

Translate your text for free

Over 30 languages available including English, French, Spanish, German, Italian, Chinese, Hindi and many others.

English ⇄ French

i love you

Je t'aime

10 characters / 6000

je

I

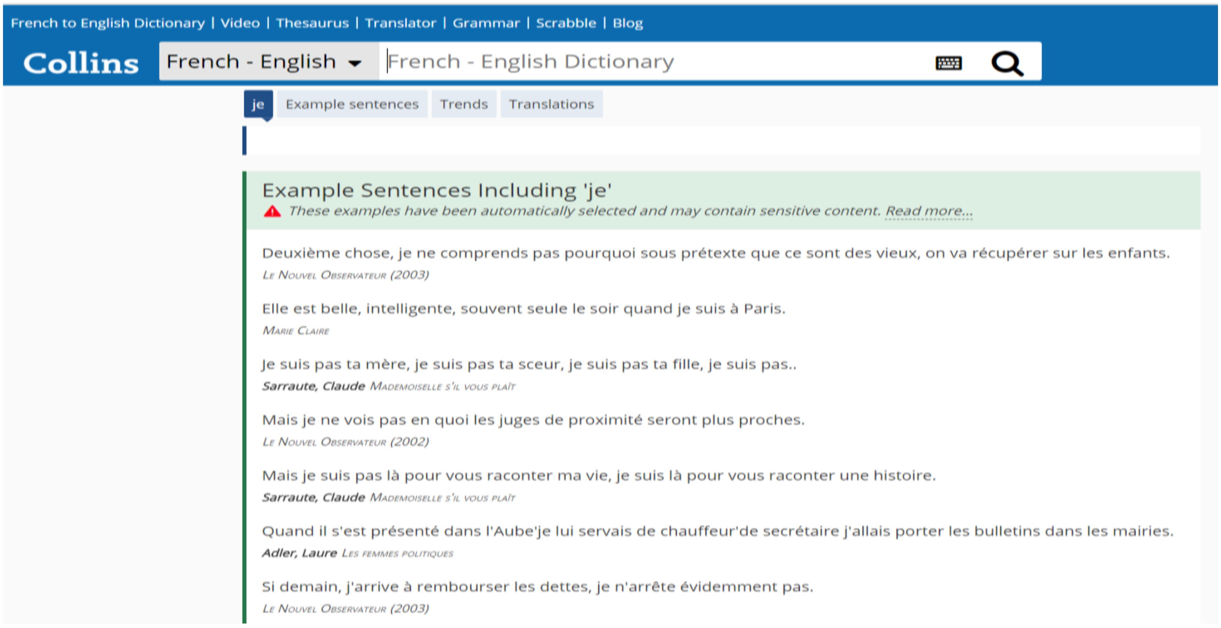
⇒ Je t'appellerai ce soir. → I'll phone you this evening.

⇒ J'arrive! → I'm coming!

⇒ J'hésite. → I'm not sure.

Figure 6.3.3: Treatment of the phrase 'I love you' in Collins dictionary https://www.collinsdictionary.com/dictionary/englishfrench

When the user clicks on either one of the examples of *Je*, they are taken into another window which gives them other examples of usage. See the following illustration taken from the dictionary. So, it has examples that utilize *je* in their sentence structures. This way, the user can better practice and get used to using *je* in many forms.


French to English Dictionary | Video | Thesaurus | Translator | Grammar | Scrabble | Blog
Collins
French - English
French - English Dictionary
je
Example sentences
Trends
Translations

Example Sentences Including 'je'

⚠ *These examples have been automatically selected and may contain sensitive content. Read more...*

Deuxième chose, je ne comprends pas pourquoi sous prétexte que ce sont des vieux, on va récupérer sur les enfants.
Le Nouvel Observateur (2003)

Elle est belle, intelligente, souvent seule le soir quand je suis à Paris.
Marie Claire

Je suis pas ta mère, je suis pas ta sceur, je suis pas ta fille, je suis pas..
Sarraute, Claude *Mademoiselle s'il vous plaît*

Mais je ne vois pas en quoi les juges de proximité seront plus proches.
Le Nouvel Observateur (2002)

Mais je suis pas là pour vous raconter ma vie, je suis là pour vous raconter une histoire.
Sarraute, Claude *Mademoiselle s'il vous plaît*

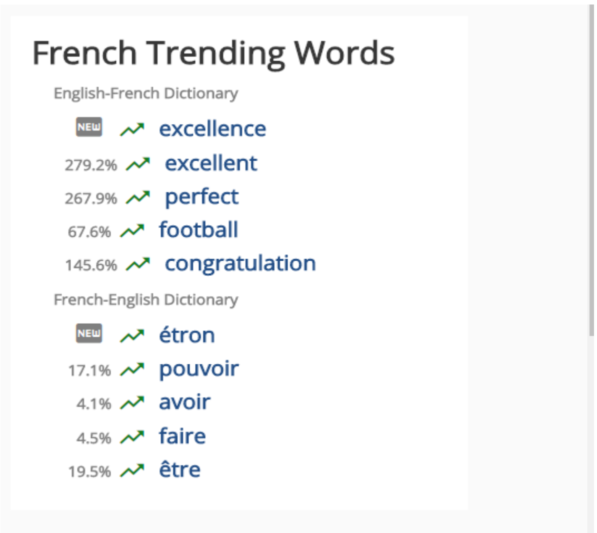
Quand il s'est présenté dans l'Aube'je lui servais de chauffeur'de secrétaire j'allais porter les bulletins dans les mairies.
Adler, Laure *Les femmes politiques*

Si demain, j'arrive à rembourser les dettes, je n'arrête évidemment pas.
Le Nouvel Observateur (2003)

Figure 6.3.4: Treatment of *Je* ‘I’ in Collins online dictionary (https://www.collinsdictionary.com/dictionary/english-french)

This treatment of *je* is good, especially for the user. However, one cannot but wonder why they treated the subject pronoun *I (je)* and totally disregarded the other words that appear with it. This is a real cause for concern. Especially because space is not a problem when it comes to online dictionaries. Hence De Schryver (2003:2) posits that: “The wildest futuristic dreams revolve around multimedia Internet dictionaries, for which space restrictions disappear, and for which the output can be tailored to suit each unique user”. Rundell (2015:318) shares this view, saying that “without the space constraints imposed by the printed medium, publishers of online dictionaries are experimenting with new ways of providing larger numbers of examples”.

The following discussion focuses on the third box, which shows trending words in both French and English. Trending words are words that have been looked for by many users in a certain period of time and are the most frequently searched for. The words could trend for a day, week, month or year. The dictionary does not say how it selects words that are trending nor does it show how it updates them. However, the words on the trend list may be removed or added, depending on how popular they have been in a given time, as stipulated by the compilers of the dictionary. They also appear to be regularly updated even though the dictionary does not say or show, by dates, how often. Through observation, the compilers of this dictionary decided to show trending words in each of the languages over time. They use upward, downward and linear arrows with different colours and percentages on the side, which go hand in hand with the arrows. Red stands for words that are no longer trending in that particular time-frame (which only the compilers know). Green represents linear and upward arrows. The former is for words that did not move on that particular time frame, while the latter means that the words are increasingly being searched for by users. Consider the examples in figures 6.3.5 and 6.3.6 where the different codes are explained. Please note that these illustrations do not show the linear arrow explained above.



French Trending Words

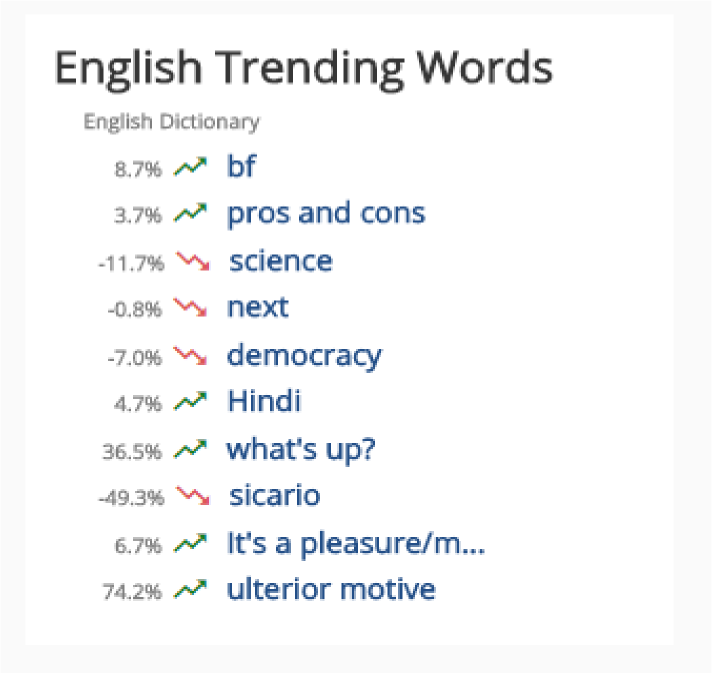
English-French Dictionary

NEW	↗	excellence
279.2%	↗	excellent
267.9%	↗	perfect
67.6%	↗	football
145.6%	↗	congratulation

French-English Dictionary

NEW	↗	étron
17.1%	↗	pouvoir
4.1%	↗	avoir
4.5%	↗	faire
19.5%	↗	être

Figure 6.3.5: French trending words in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)



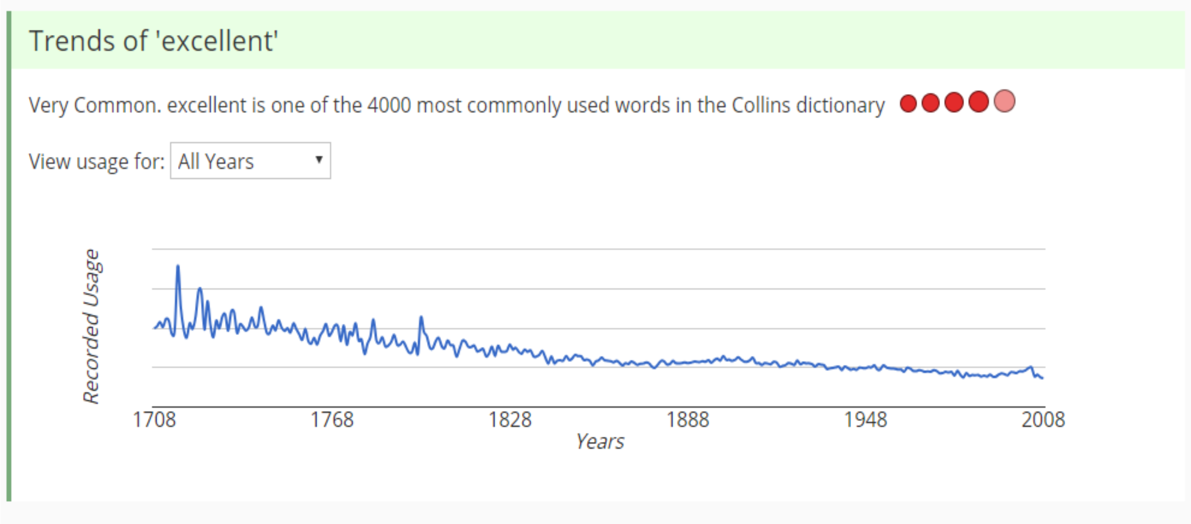
English Trending Words

English Dictionary

8.7%	↗	bf
3.7%	↗	pros and cons
-11.7%	↘	science
-0.8%	↘	next
-7.0%	↘	democracy
4.7%	↗	Hindi
36.5%	↗	what's up?
-49.3%	↘	sicario
6.7%	↗	It's a pleasure/m...
74.2%	↗	ulterior motive

Figure 6.3.6: English trending words in Collins dictionary **(**https://www.collinsdictionary.com/dictionary/english-french)

The changes in patterns will be seen when users’ choices change. Looking at the example of the word *excellent*, the Collins dictionary states that it is one of the 4,000 most commonly used words in their dictionary. Its highest recorded use was in the 1700s and, from the 1800s until 2008, it has been moving in a linear direction with not much change occurring in its use. This could be caused by the coining of new words that also have the same meaning as *excellent.* These results show that the compilers of this online dictionary used the corpus to understand words that trend.



Trends of 'excellent'

Very Common. excellent is one of the 4000 most commonly used words in the Collins dictionary ●●●●●

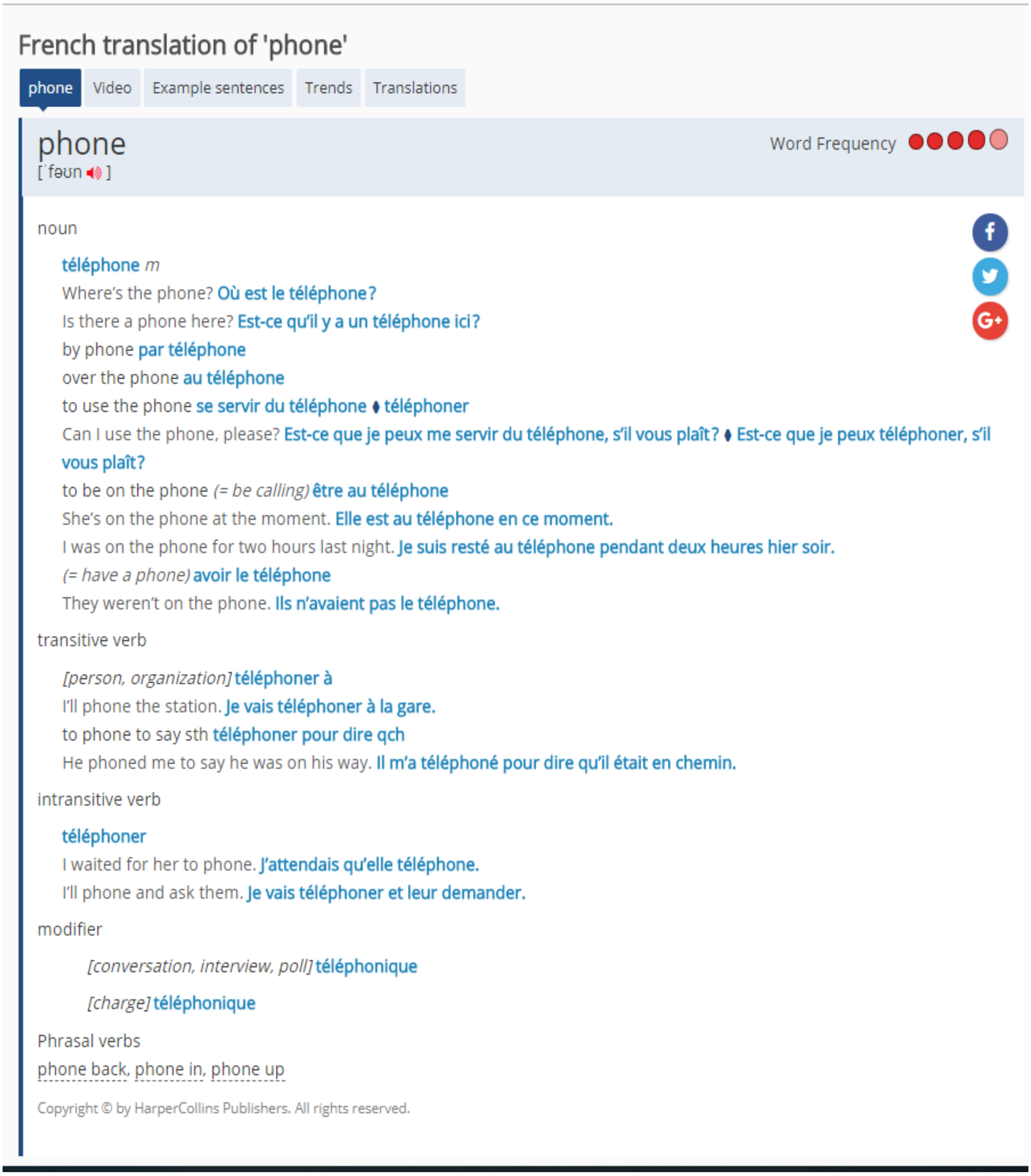
View usage for: All Years ▾

Figure 6.3.7: Word trends over the years in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)

The following illustration (figure 6.3.8) shows how the word *phone* has been treated in the Collins Online Dictionary. It is a common word and one of the 4,000 most frequently used words in the Collins dictionary. The word *phone* is explained, and examples of usage given. The dictionary shows that it is a noun first and foremost, and examples of its usage as a noun are given in English together with their French counterparts. It has been used in statements and questions, to give the user a clear indication of the noun usage it assumes. See the following example:

Where’s the phone? *Où est le téléphone?*
She’s on the phone at the moment. *Elle est au téléphone en ce moment.*

The word *telephone* is also treated in its transitive, intransitive verb forms and modifier modes with examples of usage.



French translation of 'phone'

phone
[ˈfəʊn]

Word Frequency ●●●●●

noun

télèphone *m*
Where's the phone? Où est le téléphone?
Is there a phone here? Est-ce qu'il y a un téléphone ici?
by phone par téléphone
over the phone au téléphone
to use the phone se servir du téléphone ♦ téléphoner
Can I use the phone, please? Est-ce que je peux me servir du téléphone, s'il vous plaît? ♦ Est-ce que je peux téléphoner, s'il vous plaît?
to be on the phone *(= be calling)* être au téléphone
She's on the phone at the moment. Elle est au téléphone en ce moment.
I was on the phone for two hours last night. Je suis resté au téléphone pendant deux heures hier soir.
(= have a phone) avoir le téléphone
They weren't on the phone. Ils n'avaient pas le téléphone.

transitive verb

[person, organization] téléphoner à
I'll phone the station. Je vais téléphoner à la gare.
to phone to say sth téléphoner pour dire qch
He phoned me to say he was on his way. Il m'a téléphoné pour dire qu'il était en chemin.

intransitive verb

téléphoner
I waited for her to phone. J'attendais qu'elle téléphone.
I'll phone and ask them. Je vais téléphoner et leur demander.

modifier

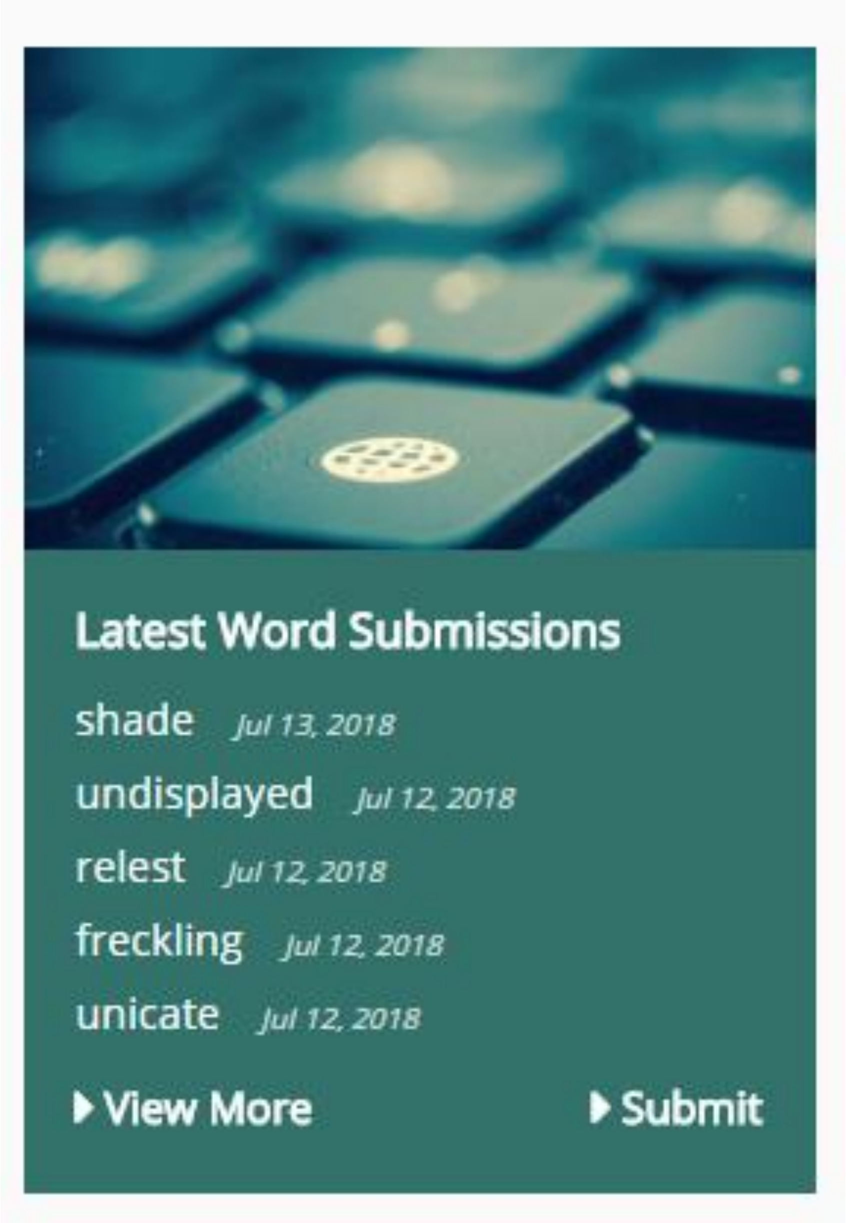
[conversation, interview, poll] téléphonique
[charge] téléphonique

Phrasal verbs
phone back, phone in, phone up

Copyright © by HarperCollins Publishers. All rights reserved.

Figure 6.3.8: Treatment of phone in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)

Users submit newly coined or used words that will be investigated by the relevant people before they are put in the dictionary. Therefore, lexicographers do not have to be out and about, trying to find the latest words and their examples of usage – because many words likely to be looked for are provided through submissions by other users. As a contributor, all that you have to do is to sign up for an account, log on, and submit, then wait for investigations to be done. Lexicographers are also at an advantage here because they get the words from people on the ground who use them on a daily basis.



Latest Word Submissions

shade Jul 13, 2018

undisplayed Jul 12, 2018

relest Jul 12, 2018

freckling Jul 12, 2018

unicate Jul 12, 2018

Figure 6.3.9: Word Submissions in Collins dictionary (https://www.collinsdictionary.com/dictionary/english-french)

6.4 Analysis of the Macmillan English Dictionary (MED)

The Macmillan English Dictionary decided in 2013 that it would discontinue its printed dictionaries and rather focus only on the modern and innovative online dictionaries, which provides endless possibilities and has taken dictionaries to new heights. The decision could not have been an easy one, or without criticisms from those who either never believed in the potential of dictionaries in a digital world or those who were scared to change to this new medium. Hence the Editor-in-Chief of Macmillan, Rundell, M. tells the critics of their online dictionary to just try it on a video shared on *YouTube* titled: *Macmillan Dictionary: Our move from print to online your questions answered*.

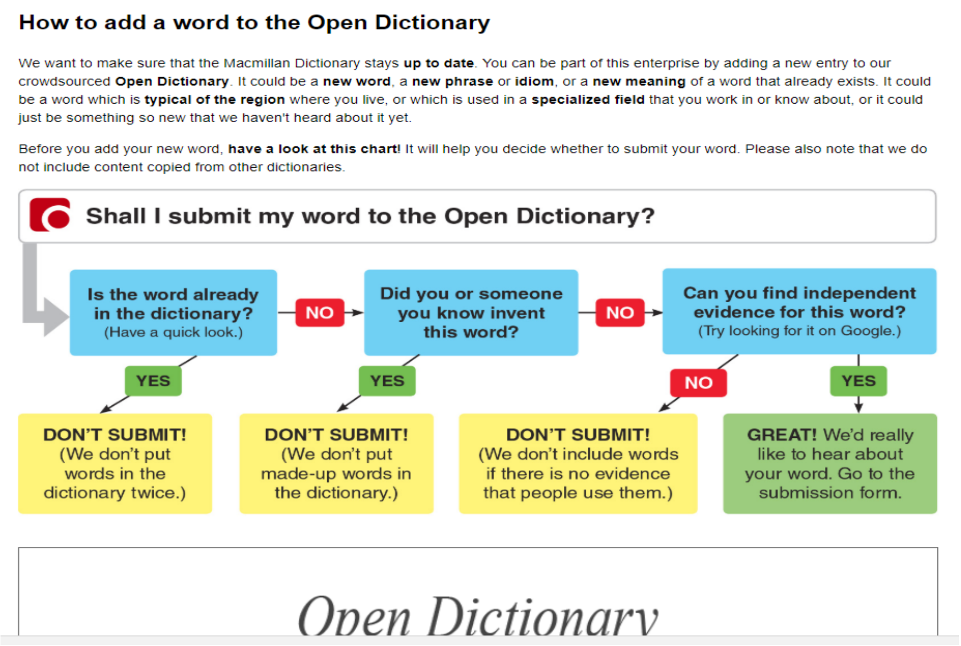
As Macmillan Education (n.d.) reported:

> “With this migration to new media Rundell believes that Macmillan's dictionaries that Macmillan's **dictionaries have found their ideal medium**: The traditional book format is very limiting for any kind of reference work. Books are out of date as soon as they are printed, and the space constraints they impose often compromise our goals of clarity and completeness. There is so much more we can do for our users in digital media… [The] Macmillan Dictionary Online provides an **English dictionary and thesaurus**, as well as a popular **blog** about topical issues such as the use of *pleb* or *omnishambles*…and the **crowd-sourced 'Open Dictionary'…” Stephen Bullon**, Macmillan Education's Publisher for Dictionaries also supports his colleague by noting that “their research reveals that most people today get their reference information from either their computer, tablet, or phone, and continued that the message is clear and unambiguous: **the future of the dictionary is digital**.”

> “…with unlimited space and digital functions such as multimedia and hyperlinking, new media provide exciting opportunities for innovation and improved coverage and open up endless possibilities for reference resources which will serve users’ needs more effectively than their print bound predecessors.” (Rundell 2015:303)

Delving into the MacMillan English Dictionary itself, the word *house* is used as an example to see how this particular dictionary treats words. Following is a screen shot (figure 6.4.1) of the treatment of *house* and what captured my attention firstly was the 'show more or less' icon on the far right, which a user can use depending on the type of information they want to see. The dictionary provides the user with both the definition and synonyms of words they are searching for, thereby making this dictionary a great source of information and tool for learning. It also offers the reader words related to their search. Like the Collins Online Dictionary, Macmillan also has a platform where ordinary users can contribute towards increasing words and meanings and say they "want to make sure that the Macmillan Dictionary stays up to date".

The most commendable thing about Macmillan Dictionary is that they give the contributor instructions on how to contribute, giving them notes on what to do and not do if they are interested in submitting words. (As seen below in figure 6.4.1)



How to add a word to the Open Dictionary

We want to make sure that the Macmillan Dictionary stays **up to date**. You can be part of this enterprise by adding a new entry to our crowdsourced **Open Dictionary**. It could be a **new word**, a **new phrase** or **idiom**, or a **new meaning** of a word that already exists. It could be a word which is **typical of the region** where you live, or which is used in a **specialized field** that you work in or know about, or it could just be something so new that we haven't heard about it yet.

Before you add your new word, **have a look at this chart!** It will help you decide whether to submit your word. Please also note that we do not include content copied from other dictionaries.

Shall I submit my word to the Open Dictionary?

Is the word already in the dictionary? (Have a quick look.)

NO

Did you or someone you know invent this word?

NO

Can you find independent evidence for this word? (Try looking for it on Google.)

YES

DON'T SUBMIT! (We don't put words in the dictionary twice.)

YES

DON'T SUBMIT! (We don't put made-up words in the dictionary.)

NO

DON'T SUBMIT! (We don't include words if there is no evidence that people use them.)

YES

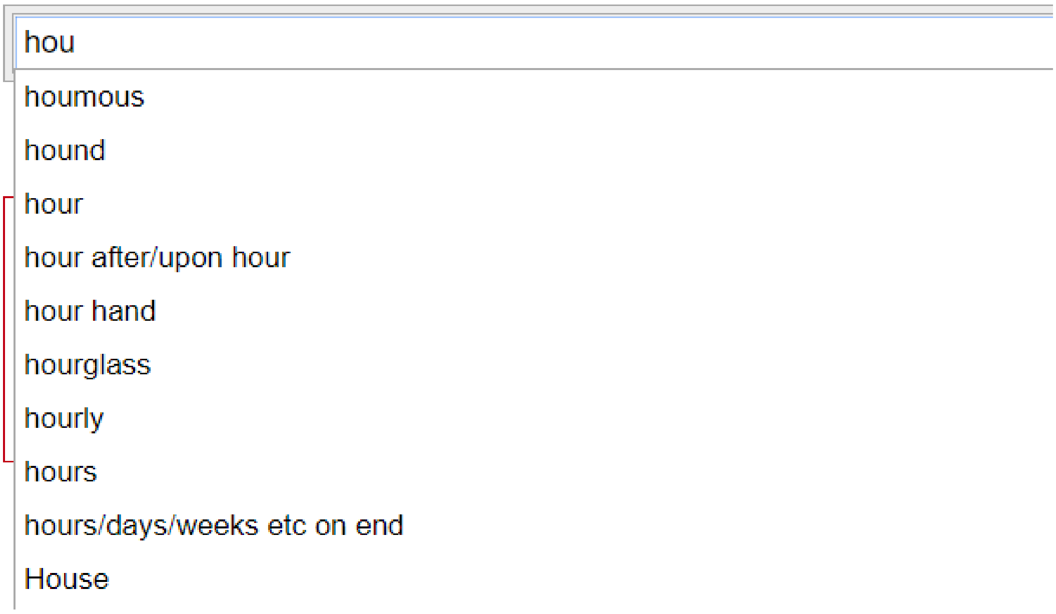
GREAT! We'd really like to hear about your word. Go to the submission form.

Open Dictionary

Figure 6.4.1: Addition of words to the Open Dictionary in Macmillan dictionary (https://www.macmillandictionary.com/open-dictionary/submit.html)

This dictionary also has a pronunciation icon. It also has the Word Forms icon that instantly gives the singularity or plurality of words. The list under Menu comprises secondary meanings of the word *house*. In this case there are about eight (8) meanings of *house*, which when clicked on take the user to the correct context. It also informs the user of other possibilities for the use of this particular word.

When searching for a word, the user does not have to type in the whole word but is provided with possibilities, based on the alphabet or phonetics. For instance, in searching for *house* one only needs to type *hou* – and every word that begins like that is automatically retrieved by their software. As shown in the illustration below, (figure 6.4.2) the words appear in their particular order. The word *house* is right at the end. The scenario may be different for other words, as some may appear among the first few or in the middle. They are clickable and, once clicked on, they take you to the meaning of the word.



hou

houmous

hound

hour

hour after/upon hour

hour hand

hourglass

hourly

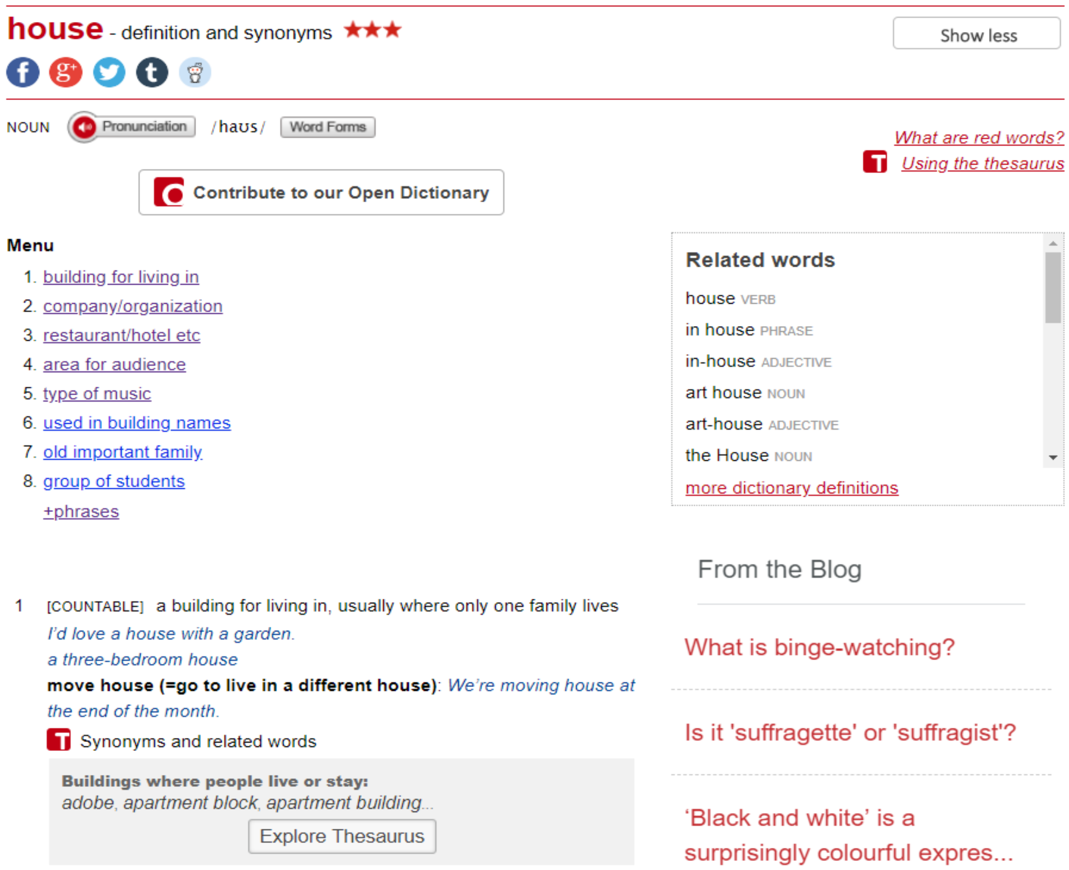
hours

hours/days/weeks etc on end

House

Figure 6.4.2: Treatment of hou-

https://www.macmillandictionary.com



house - definition and synonyms ★★★

NOUN /haʊs/

Menu

1. building for living in
2. company/organization
3. restaurant/hotel etc
4. area for audience
5. type of music
6. used in building names
7. old important family
8. group of students

+phrases

1 [COUNTABLE] a building for living in, usually where only one family lives

I'd love a house with a garden.

a three-bedroom house

move house (=go to live in a different house): *We're moving house at the end of the month.*

Synonyms and related words

Buildings where people live or stay:

adobe, apartment block, apartment building...

Related words

house VERB

in house PHRASE

in-house ADJECTIVE

art house NOUN

art-house ADJECTIVE

the House NOUN

more dictionary definitions

From the Blog

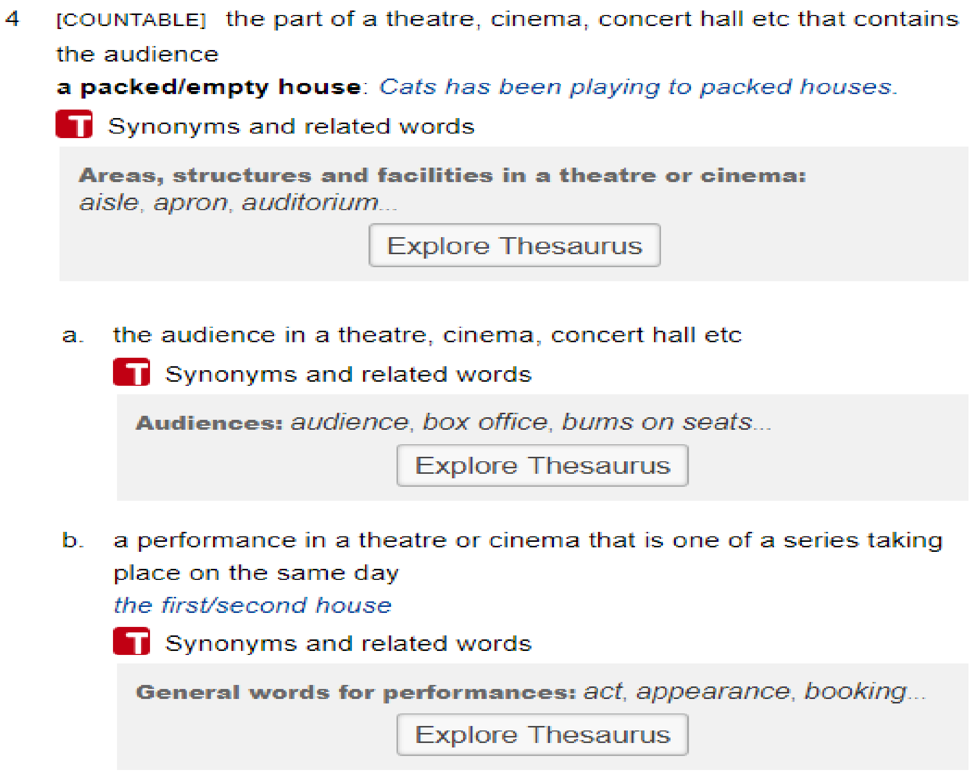
What is binge-watching?

Is it 'suffragette' or 'suffragist'?

'Black and white' is a surprisingly colourful expres...

Figure 6.4.3: Treatment of house (https://www.macmillandictionary.com)

For example, if a user clicks on number 4, *area for audience*, he/she is immediately taken to the meaning they want, which will be illustrated in the following diagram (figure 6.4.4).



4 [COUNTABLE] the part of a theatre, cinema, concert hall etc that contains the audience

a packed/empty house: *Cats has been playing to packed houses.*

T Synonyms and related words

Areas, structures and facilities in a theatre or cinema: *aisle, apron, auditorium...*

a. the audience in a theatre, cinema, concert hall etc

T Synonyms and related words

Audiences: *audience, box office, bums on seats...*

b. a performance in a theatre or cinema that is one of a series taking place on the same day

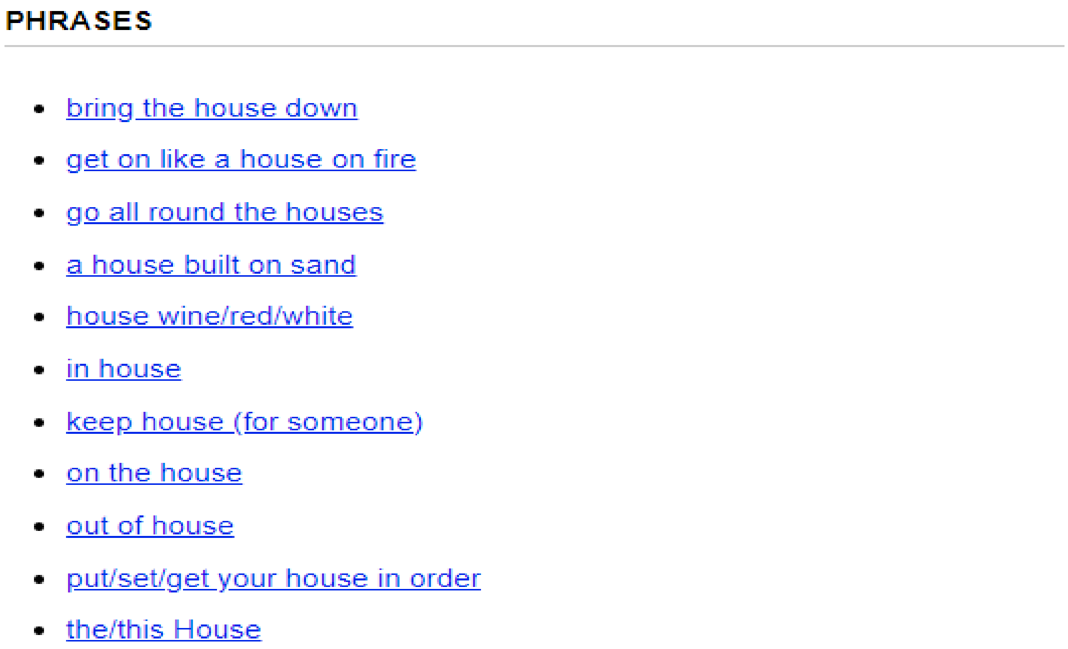
the first/second house

T Synonyms and related words

General words for performances: *act, appearance, booking...*

Figure 6.4.4: Illustration of meaning of house under 'area of audience' **(**https://www.macmillandictionary.com**)**

The user is also given a choice or option to explore the thesaurus if they want to. They are also provided with examples of use which are helpful to any learner of the English language. Under the number 8 there is an icon +*phrases* (see figure 6.4.5). When the user clicks on it, they are provided with the next screen which consists of English phrases that relate to *house*.

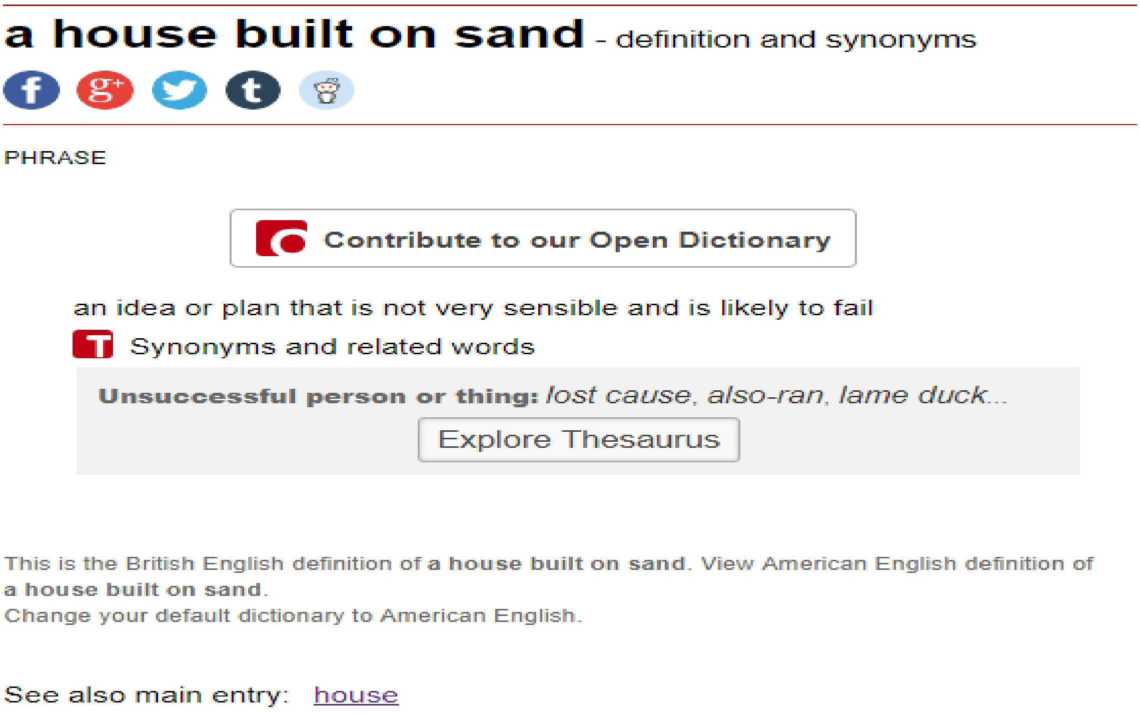


PHRASES

- bring the house down
- get on like a house on fire
- go all round the houses
- a house built on sand
- house wine/red/white
- in house
- keep house (for someone)
- on the house
- out of house
- put/set/get your house in order
- the/this House

Figure 6.4.5: Phrases that relate to house (https://www.macmillandictionary.com)

When the user clicks on *a house built on sand* they are automatically taken to the meaning of that phrase together with its synonyms and related words.



a house built on sand - definition and synonyms

PHRASE

an idea or plan that is not very sensible and is likely to fail

Synonyms and related words

Unsuccessful person or thing: *lost cause, also-ran, lame duck...*

This is the British English definition of **a house built on sand**. View American English definition of **a house built on sand**.
Change your default dictionary to American English.

See also main entry: house

Figure 6.4.6: Example of meaning of phrase (https://www.macmillandictionary.com)

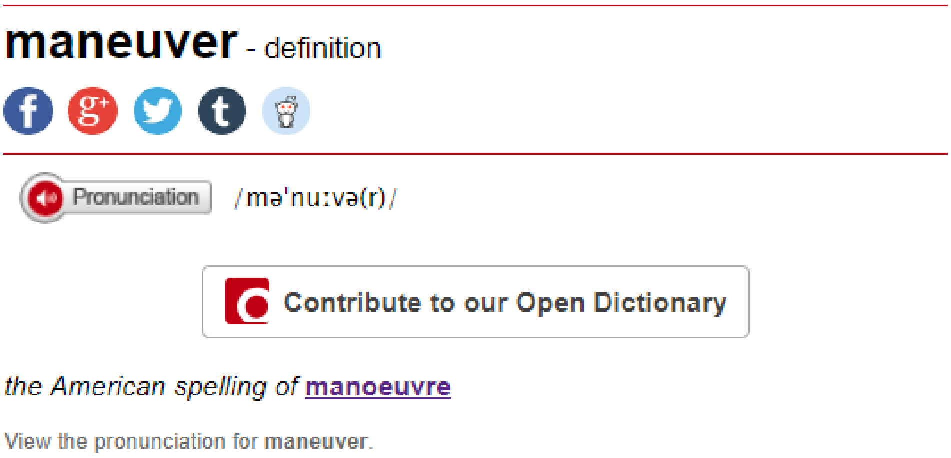
Another interesting aspect is that the dictionary gives the user a choice between a definition in British English and American English, and they are also given an option to change their default dictionary settings to American English. "When MED is launched it immediately opens up on a random lemma which is automatically pronounced in British English and clickable options for both British and American English are provided." (Prinsloo 2005:12).

The following relevant features listed by Prinsloo (2005:12) give electronic dictionaries a huge advantage in terms of ease of access and putting the user at ease, because they encourage the user to manoeuvre around the dictionary. These features are:

a. Pop-up access
b. Bringing together of related items
c. New routes to the data
d. Less dependency on alphabetical order

e. Fuzzy spelling (overriding the user's mistakes/ even if it is spelled wrong, the dictionary tries to find the word)
f. Intelligent extrapolation of characters keyed in
g. Audible pronunciation

However, quite interestingly, when an (incorrectly spelled) word like *maneuver* is typed in, the dictionary recognises that word – but instead of giving the definition, it takes the user to the American spelling.



maneuver - definition

Pronunciation /mə'nuːvə(r)/

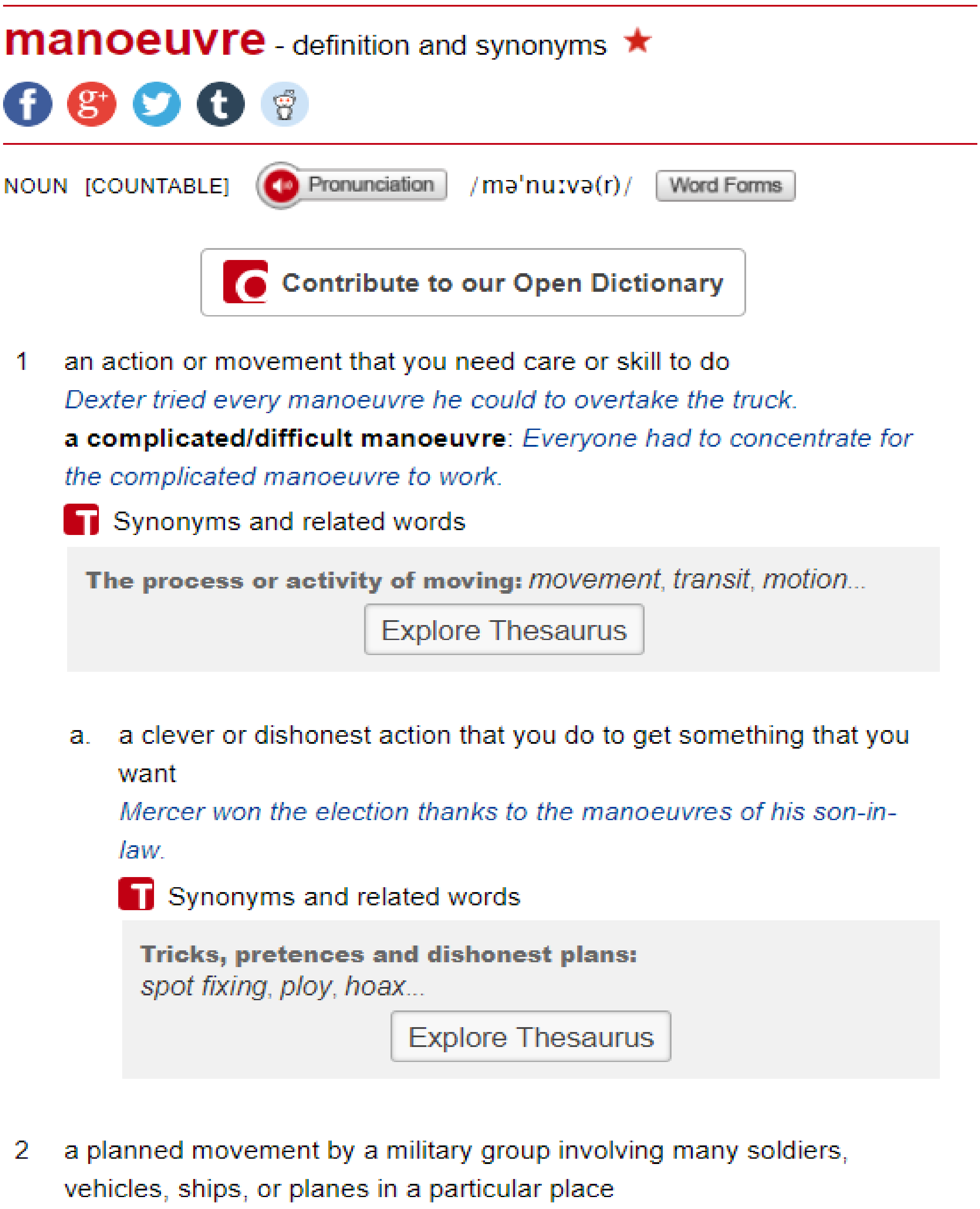
Contribute to our Open Dictionary

the American spelling of **manoeuvre**

View the pronunciation for **maneuver**.

Figure 6.4.7: Treatment of 'maneuver' in British English (https://www.macmillandictionary.com/dictionary/british/manoeuvre_1)

It is only when the user clicks on the American spelling that they are able to get a full treatment of the word. It seems as if the user is forced to use the American spelling over the British one. See below illustration.



manoeuvre - definition and synonyms ★

NOUN [COUNTABLE] /məˈnuːvə(r)/

1 an action or movement that you need care or skill to do

Dexter tried every manoeuvre he could to overtake the truck.

a complicated/difficult manoeuvre: *Everyone had to concentrate for the complicated manoeuvre to work.*

Synonyms and related words

The process or activity of moving: *movement, transit, motion...*

a. a clever or dishonest action that you do to get something that you want

Mercer won the election thanks to the manoeuvres of his son-in-law.

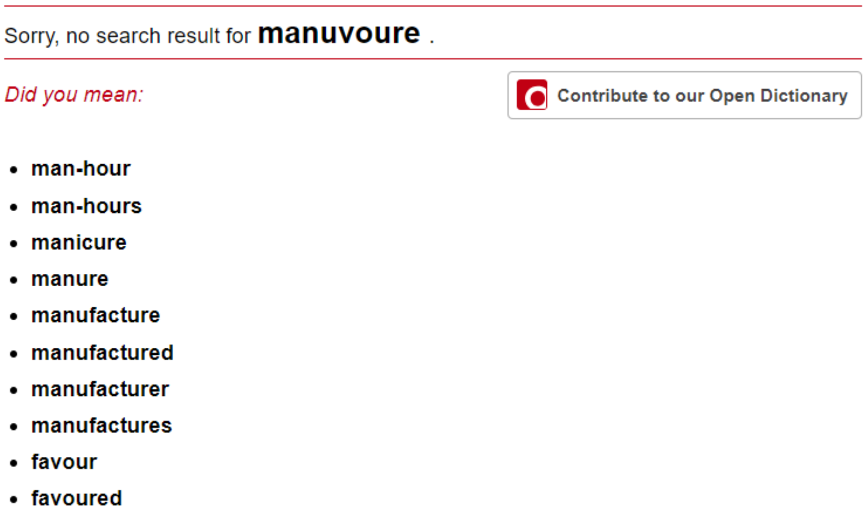
Synonyms and related words

Tricks, pretences and dishonest plans:
spot fixing, ploy, hoax...

2 a planned movement by a military group involving many soldiers, vehicles, ships, or planes in a particular place

Figure 6.4.8: Treatment of 'manoeuvre' in American spelling (https://www.macmillandictionary.com/dictionary/british/manoeuvre_1)

The issue of fuzzy spelling is another point which is not dealt with properly. The dictionary software is not able to pick up what the user wanted to write, or rather the mistake they made. Going back to the example of *maneuver/manoeuvre*, if a user does not know the correct spelling, they will not get what they are looking for from the dictionary. They will simply get a *'sorry, no search result for manuvoure'* (see figure 6.4.9). This idea means that a dictionary user needs to have certain language skills for them to fully comprehend and access the dictionary to its full potential.



Sorry, no search result for **manuvoure** .

Did you mean:

- **man-hour**
- **man-hours**
- **manicure**
- **manure**
- **manufacture**
- **manufactured**
- **manufacturer**
- **manufactures**
- **favour**
- **favoured**

Figure 6.4.9: Example of wrong spelling of manoeuvre (https://www.macmillandictionary.com)

6.5 Conclusion

This chapter provided an in-depth analysis and examination of the free Collins English-French online dictionary and the free Macmillan English Dictionary (MED). Online dictionaries have many advantages, including the ability for users to have access to dictionaries without having to carry a heavy book around or having to page through many pages. However, their downside is their constant need for internet connectivity for them to be accessed. Thus, in as much as online dictionaries have their pros and cons, which have been discussed in this chapter, they have revolutionized the dictionary-making process in general. With the many features that come with them, the work of the lexicographer is more rewarding, e.g., finding new words that are used by people on the ground because they are now able to add those words themselves. On the part of the user, the availability of the pronunciation icon is an excellent innovation for new learners of languages, and it is available in both the Collins dictionary and the MED.

Chapter 7: Critical analysis of existing Sesotho dictionaries

7.1 Introduction

This chapter will explore one monolingual dictionary (Sethantšo sa Sesotho), one bilingual (Southern Sotho-English Dictionary) and one internet dictionary (Bukantswe Online). Dictionaries for African languages are often regarded as not of a high lexicographic standard and there are very few of them available in the market. A Euro-centric approach to dictionary compilation is often cited as one of the reasons for the poor quality in African language dictionaries, as Prinsloo (2015) explains.

> More than a decade ago De Schryver (2003:2) found that there were already nearly 200 internet dictionaries for nearly 120 different African languages at the time. However, from a lexicographic perspective most of these sources were merely word lists or lemma lists with only basic information on form and meaning, or paper dictionaries presented online. They do not answer the expectation of being dictionaries with true electronic features, appealing and effective screen presentation and the ultimate: online dictionaries solving lexicographic problems that could not be satisfactorily solved in a paper dictionary.

Awak (1990:17) notes that:

> The history of lexicography in Africa began as a result of European activities: exploration, evangelization and colonialization. The early lexicons, whether compiled by explorers, missionaries or colonial administrators, were 'Euro-centred', produced in Europe for Europeans rather than for African users … Even with the emergence of modern linguistics, lexicographic works have been primarily intended for scholarly interest and not for the needs of ordinary Africans.

Khumalo (2015:21) reflects negatively on attitudes towards African languages in general by stating that African languages have been disregarded, discredited, minimized and associated with negativity in the broader context, hence the development of African lexicography and dictionary compilation has been slow.

> The use of African languages has hitherto been discredited. Their use is despised, discouraged and even feared by prominent government officials who should know better because they are viewed wrongly as divisive and of limited vernacular use in our society with a glaring lack of capacity to contribute meaningfully to the knowledge economy. (Khumalo 2015:21)

As much as Africa is home to more than 2,000 languages, the sad reality faced by many of these languages is extinction if efforts are not made to recapture and preserve them. This sad reality is reiterated by the United Nations Educational, Scientific and Cultural Organization (UNESCO) in the following statement:

> It is estimated that, if nothing is done, half of the over 6,000 plus languages spoken today will disappear by the end of this century. With the disappearance of unwritten and undocumented languages, humanity would lose not only an in irreplaceable cultural heritage, but also valuable ancestral knowledge embedded, in particular, in indigenous languages. However, this process is neither inevitable nor irreversible: well-planned and effectively implemented language policies can bolster ongoing efforts of speaker communities to maintain or revitalize their mother tongues and pass them on to younger generations. (UNESCO, n.d.)

Khumalo (2015:22-23) expresses a similar concern:

> For African languages to be able to be used successfully as media of instruction an incredible amount of resolve and commitment is needed from African governments, educationists, social and natural scientists together with language experts.

These views all point to the fact that African languages are not used enough in formal settings, like teaching and training institutions, the economic sector, in scientific studies, etc. The effects of this have been detrimental to the country. It is fair to conclude that the nation’s young people are the real victims of this situation. This should make us worry about the future of South Africa if the most critical challenges are not dealt with properly.

Khumalo (2015:22) notes that Africa is one of the few continents in the world where children receive knowledge in foreign tongues and, according to numerous scholars, this is the main reason why many children in Africa drop out of school and why there is such a high failure rate of African learners. Thus, if culture is the main determinant of people's behaviour, then language is the central feature of culture. It is the medium through which culture is transmitted, interpreted and configured, and it plays an important role – to the extent that some scientists believe that, without language, culture would not be possible because language simultaneously reflects culture and is influenced by it. (Jiang 2000:382)

Secondly and most importantly, this background information is essential because it forms the core of why African languages have fewer dictionaries, and the reasons why even those that are available are not compiled according to modern methodologies. Yes, dictionaries for African indigenous languages have been compiled, but they have the following shortcomings: standard of research and quality is very poor, as a result they are not user-friendly; and the lexicographic processes involved in their development are not planned properly and not followed to the letter; entries are added in an ad hoc manner, with some words not included and lemmatised; and, for some inexplicable reasons, some words are treated differently to others in the same dictionary, etc. Corpora have a positive impact on the compilation of ideal dictionaries for the indigenous languages and if used to its full effect, it eliminates all these deficiencies.

This chapter will also show how Sesotho dictionaries lack the capacity and tenacity to be useful to the users and the importance of using corpora in modern lexicography.

> If African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should therefore become an absolute priority. (De Schryver and Prinsloo 2000:89)

7.2 Bukantswe Online

> The electronic dictionary, whether on CD-ROM, online, or hand-held will supersede the paper dictionary in ways unimaginable in the paper-dictionary dimension, just as the computer has completely superseded the typewriter. Many

> lexicographers are of the opinion that, generally speaking, the paper dictionary has attained its maximum potential. (Prinsloo 2001:139-140)

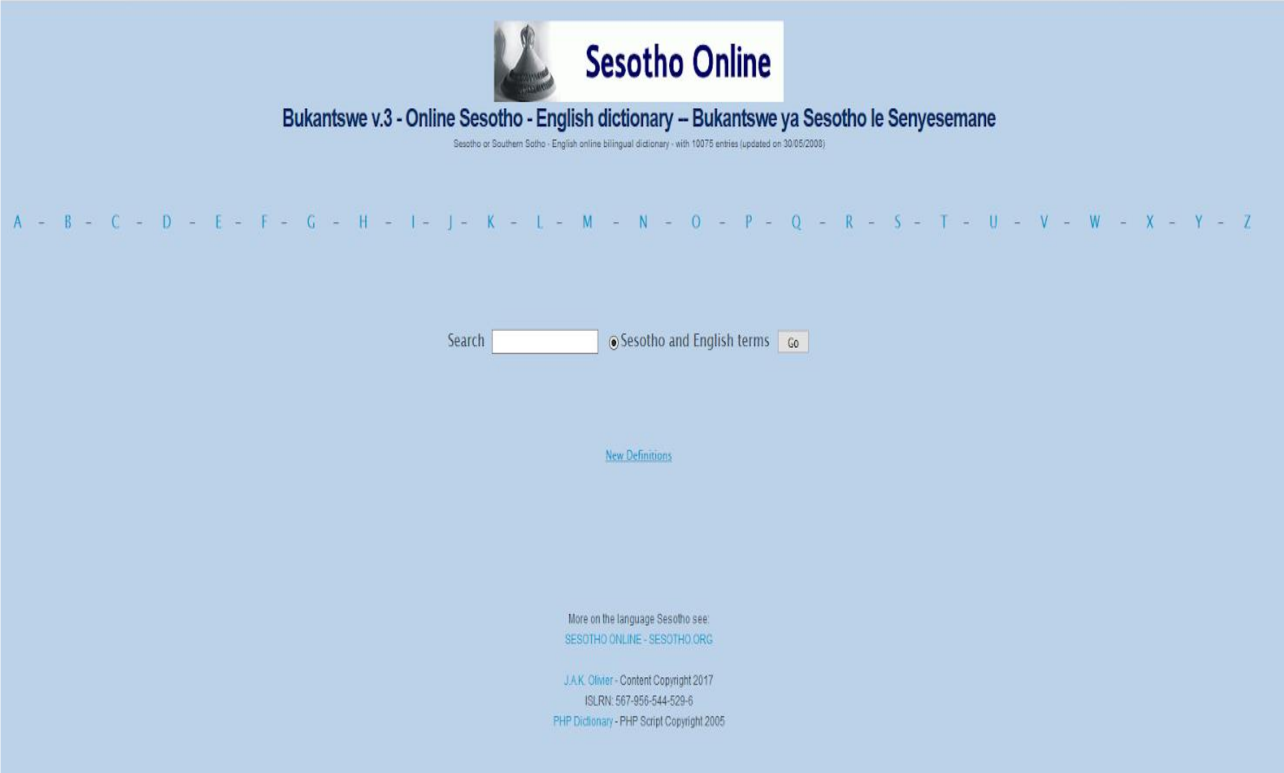
Bukantswe Online is an online dictionary that has more than 10,000 Sesotho entries that go together with their English equivalents. Searches can be done in English and Sesotho with the option of just looking for a single or any instance of the word. It is a downloadable online Sesotho-English bilingual dictionary available from http://bukantswe.sesotho.org/ However, this dictionary gives users translation equivalents of what they need and no explanatory information or additional remarks.

> When people consult a bilingual dictionary they seldom realise that the information given is not essentially a statement about meaning but a list of translation equivalents, cf. Louw (1985:53, 54). The functional status of these translation equivalents is that they may be used in certain contexts to substitute the source language item. Where the specific contexts in which translation equivalents can be used to substitute the lemma are not given as part of the lexicographical treatment, it is hardly possible that the creation of semantic equivalence can lead to the establishment of communicative equivalence. (Gouws 1996:16)

This online dictionary gives users translation equivalents of words and not any semantic, etymology or syntactic information. It does not possess the same qualities that English dictionaries have, and this shows that more work needs to be done to improve Sesotho online and paper dictionaries.

> Although bilingual dictionaries are employed as polyfunctional sources of semantic information, their main function is not a transfer of meaning. Bilingual dictionaries are aids in interlingual translations and have to focus on a treatment that enables the user to render a good and sound translation. The main aim of the dictionary should not only be the establishment of a relation of semantic equivalence between source and target language. Instead, a lexicographer has to endeavour to reach communicative equivalence. (Gouws 1996:16)

At first glance when a user opens the Bukantswe Online dictionary, they come across the blue screen, as shown in figure 7.2.1, with alphabetic stretches ranging from A-Z. When each alphabetical stretch is tabbed, another screen pops up which has ‘all the information in that stretch’. The term information is used here because everything is encompassed in each letter, be it a part of speech, idioms, phrases, etc.



Sesotho Online

Bukantswe v.3 - Online Sesotho - English dictionary – Bukantswe ya Sesotho le Senyesemane

Sesotho or Southern Sotho - English online bilingual dictionary - with 10075 entries (updated on 30/05/2008)

A - B - C - D - E - F - G - H - I - J - K - L - M - N - O - P - Q - R - S - T - U - V - W - X - Y - Z

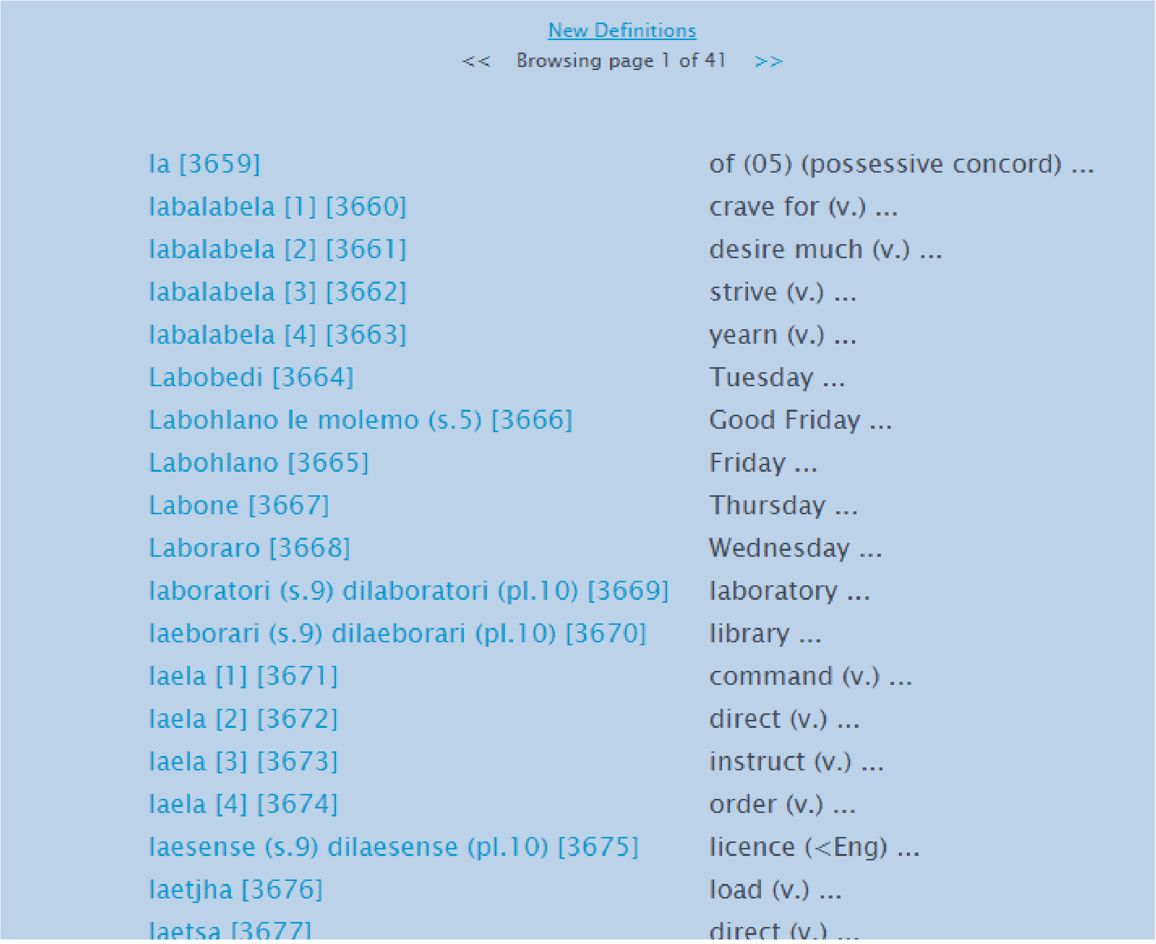
Search ● Sesotho and English terms Go

New Definitions

More on the language Sesotho see:
SESOTHO ONLINE - SESOTHO.ORG

J.A.K. Olivier - Content Copyright 2017
ISLRN: 567-956-544-529-6
PHP Dictionary - PHP Script Copyright 2005

Figure 7.2.1: A-Z sequence of alphabetical stretches from Bukantswe Online (http://bukantswe.sesotho.org/)



New Definitions

<< Browsing page 1 of 41 >>

la [3659]	of (05) (possessive concord) ...
labalabela [1] [3660]	crave for (v.) ...
labalabela [2] [3661]	desire much (v.) ...
labalabela [3] [3662]	strive (v.) ...
labalabela [4] [3663]	yearn (v.) ...
Labobedi [3664]	Tuesday ...
Labohlano le molemo (s.5) [3666]	Good Friday ...
Labohlano [3665]	Friday ...
Labone [3667]	Thursday ...
Laboraro [3668]	Wednesday ...
laboratori (s.9) dilaboratori (pl.10) [3669]	laboratory ...
laeborari (s.9) dilaeborari (pl.10) [3670]	library ...
laela [1] [3671]	command (v.) ...
laela [2] [3672]	direct (v.) ...
laela [3] [3673]	instruct (v.) ...
laela [4] [3674]	order (v.) ...
laesense (s.9) dilaesense (pl.10) [3675]	licence (<Eng) ...
laetjha [3676]	load (v.) ...
laetsa [3677]	direct (v.) ...

Figure 7.2.2: Extraction of letter 'L' from Bukantswe Online (http://bukantswe.sesotho.org/L)

The letter 'L' was tapped here, and this screen popped up (figure 7.2.2). This section has a list of words starting with the letter 'l' and it goes on and on until page 41. The user is presented with all instances where the 'L' stretch appears as a first letter. See figure 7.2,2 below as an example where the letter 'L' was clicked and all instances where this particular letter appears first is presented to the user. However, it is worth noting that in this specific list, the first dictionary article is a possessive concord of Noun class 5 and there are Verbs and Nouns too. The information here is not characterised well, everything is just mixed and not listed in any particular order. This therefore puts unnecessary strain on the user who has little or no knowledge of Sesotho. Some words have parts of speech on the side, but that is only shown in English on the text written in black. The treatment of lemmas is not the same for Sesotho and

English. There are numbers on each of the Sesotho words, which are not explained in the dictionary and in the English side they are not there, and this too will confuse the user because there is no explanation why they are there or the purpose they serve.

This dictionary does not help users with pronunciation of words, especially compared to English dictionaries which have such a mechanism to help the users understand words and how they are pronounced. De Schryver and Prinsloo (2000:310) emphasize the importance of this:

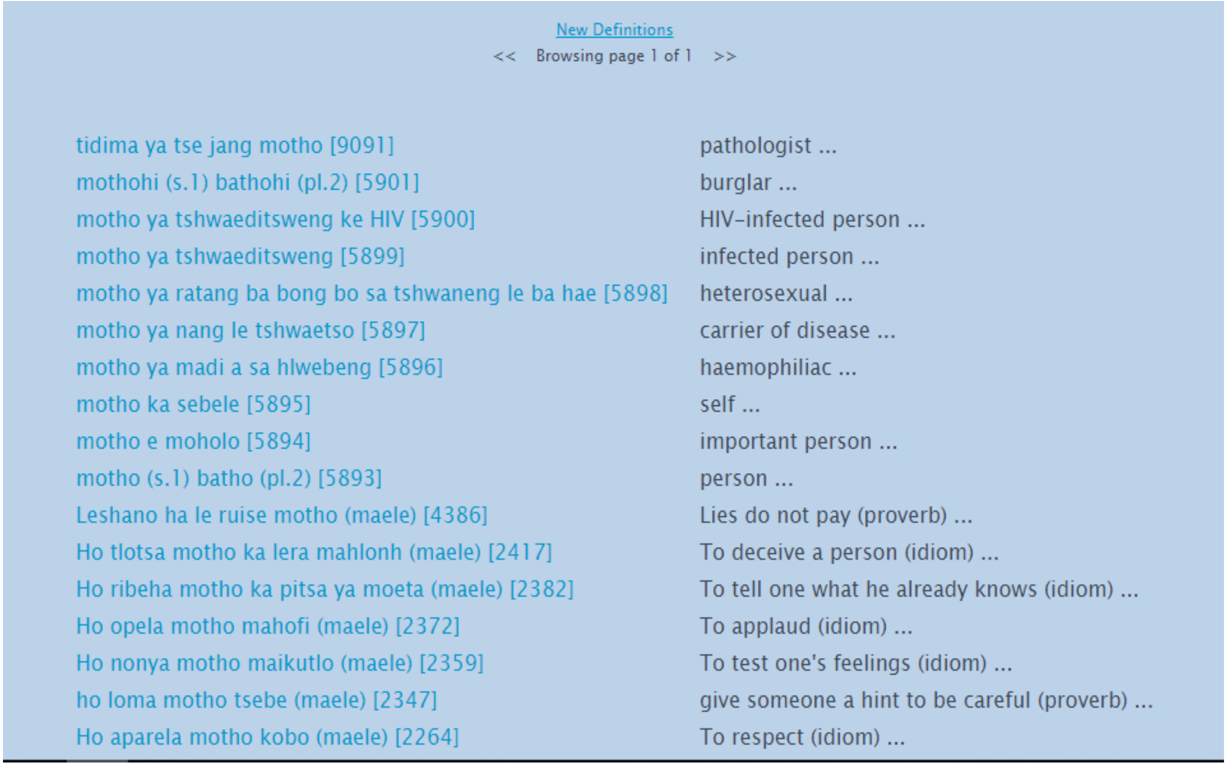
> Hence, in order to respond to user needs, modern researchers must treat real language which is most conveniently stored in electronic corpora. This current linguistic trend is especially obvious in lexicography.

7.3 Word search

The word search function of Bukantswe online dictionary is not the most efficient. It overloads users with lots of unnecessary information, for example, in figure 7.3.1, the user is trying to search for the word *motho* but instead of being directed to the word they are searching for, they are presented with a list that includes all the words that has a sound combination of *motho-*. For instance, the word *mothohi* 'burglar' in line 2 is also lemmatised and the user is presented with it first before they get to the actual *motho* 'person'. This means that a user needs to go through a lot of information before they can reach their answer. In other words, when the noun *motho* 'person' is searched, the dictionary gives the user all the instances where *motho* is present, including its occurrence in phrases, idioms, proverbs, etc. The word *motho* is assigned singularity with 's.1' and plurality with *batho* 'people' 'pl.2'. The noun class for *motho* is reflected on the singular and plural notion and the user is expected to know what those signs mean, however, there is no user guide to inform them on how to approach the dictionary and find meaning. This is one of the major flaws of this dictionary.

Users who are familiar with the grammar of Sesotho and the noun class numbering system will easily work this out. They would know that the number (1) near *motho* is an indication of the Noun Class 1 and for (2) is for Noun Class 2 respectively. But to inexperienced users these numbers must be explained. The fact that the word that is looked for is not highlighted is problematic because the user needs to go through the different words, word by word before finding the word they want. This also means that the user needs to know exactly what they are looking for or else they are going to struggle going forward and this in turn breaks the teacher-learner relationship between dictionary and user which is another flaw in this dictionary.

The dictionary does not give any synonyms where applicable, no related words, no examples of usage, no parts of speech, etc. The dictionary does not give the etymology of words, their trends or frequency like most English dictionaries would. It just gives the translation equivalent of the source language.



New Definitions

<< Browsing page 1 of 1 >>

tidima ya tse jang motho [9091]	pathologist ...
mothohi (s.1) bathohi (pl.2) [5901]	burglar ...
motho ya tshwaeditsweng ke HIV [5900]	HIV-infected person ...
motho ya tshwaeditsweng [5899]	infected person ...
motho ya ratang ba bong bo sa tshwaneng le ba hae [5898]	heterosexual ...
motho ya nang le tshwaetso [5897]	carrier of disease ...
motho ya madi a sa hlwebeng [5896]	haemophiliac ...
motho ka sebele [5895]	self ...
motho e moholo [5894]	important person ...
motho (s.1) batho (pl.2) [5893]	person ...
Leshano ha le ruise motho (maele) [4386]	Lies do not pay (proverb) ...
Ho tlotsa motho ka lera mahlonh (maele) [2417]	To deceive a person (idiom) ...
Ho ribeha motho ka pitsa ya moeta (maele) [2382]	To tell one what he already knows (idiom) ...
Ho opela motho mahofi (maele) [2372]	To applaud (idiom) ...
Ho nonya motho maikutlo (maele) [2359]	To test one's feelings (idiom) ...
ho loma motho tsebe (maele) [2347]	give someone a hint to be careful (proverb) ...
Ho aparela motho kobo (maele) [2264]	To respect (idiom) ...

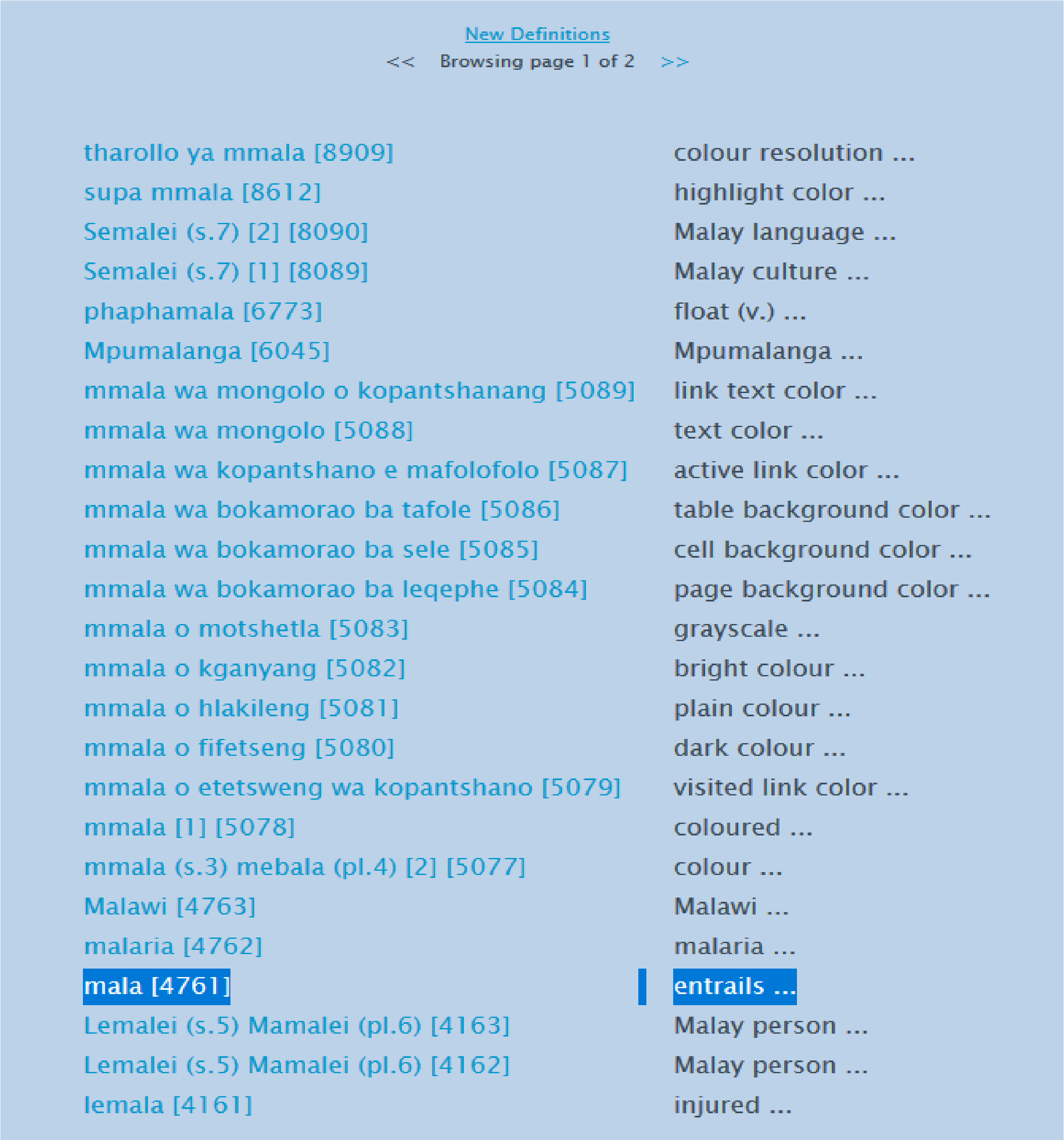
Figure 7.3.1: Extract of *motho* 'person' from Bukantswe Online (http://bukantswe.sesotho.org/display.php?action=search&word=motho&type=full)

There is a lack of dictionaries in Sesotho, hence many Basotho people end up using this dictionary as their main source. However, because the dictionary only has a limited number of words, compared to Collins dictionary which has more than 230, 000 words with translation equivalents, users are likely to be frustrated because they can't find what they were looking for. Had a corpus been used in the compilation of this dictionary, many discrepancies could have been avoided. Materials would have been collected, a corpus compiled, frequency tests made, and so on. Even though it is known that African languages have limited resources, an electronic corpus is a crucial aspect in modern linguistics.

> African corpora must be - and are being – built, hence it is beyond doubt that any first approach to corpora for the African languages cannot even come close either to the size or thoroughness that characterises today's major English corpora. Nor do we have the necessary corpus traditions, nor the necessary linguistic descriptions, nor the necessary theoretical frameworks, nor the necessary human resources, nor the necessary funds, nor the necessary demand – to name but a few - to warrant such a tremendous effort. (De Schryver and Prinsloo 2000: 96)

Furthermore, Prinsloo (2015) clearly indicates the value of even a very limited corpus in the compilation of dictionaries. It is therefore important to note that it is only through the use of corpora that a good dictionary that contains all words most likely to be looked for and one that is user-friendly can be compiled.

Consider the example highlighted below of *mala* ‘entrails’ (figure 7.3.2). The user needs to 1) go through 21 lines before reaching the translation, and 2) the user is taken through words and phrases that has *mmala* first before they reach their intended word. The alphabetic ordering is also confusing because in a normal dictionary a user would find ‘*mma*…’ after ‘*ma*…’ and not before.



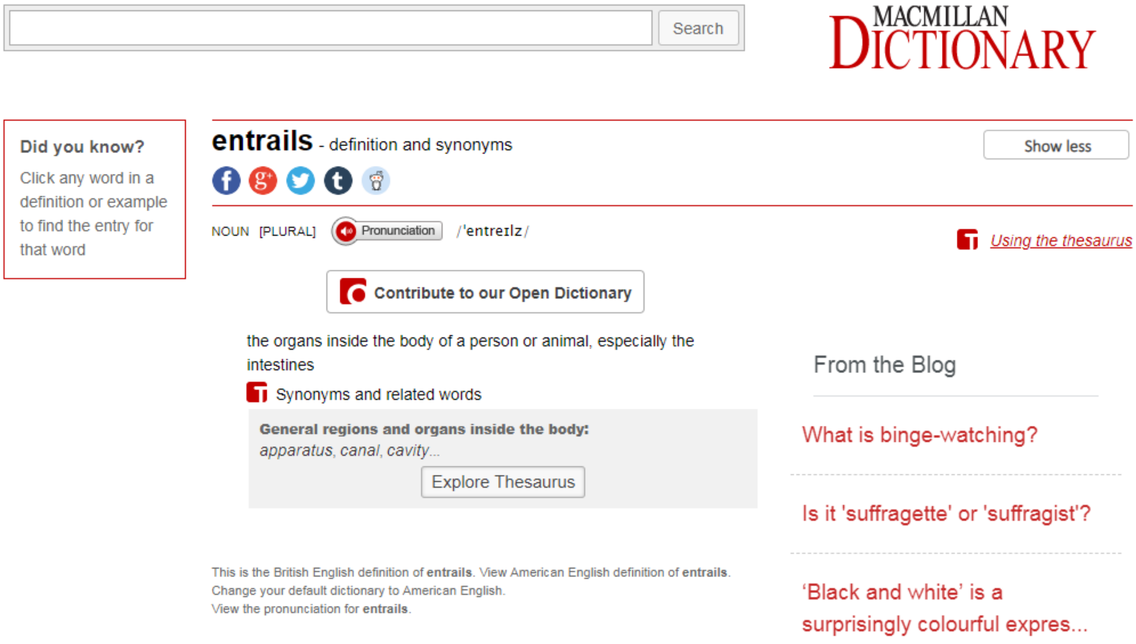
New Definitions

<< Browsing page 1 of 2 >>

tharollo ya mmala [8909]	colour resolution ...
supa mmala [8612]	highlight color ...
Semalei (s.7) [2] [8090]	Malay language ...
Semalei (s.7) [1] [8089]	Malay culture ...
phaphamala [6773]	float (v.) ...
Mpumalanga [6045]	Mpumalanga ...
mmala wa mongolo o kopantshanang [5089]	link text color ...
mmala wa mongolo [5088]	text color ...
mmala wa kopantshano e mafolofolo [5087]	active link color ...
mmala wa bokamorao ba tafole [5086]	table background color ...
mmala wa bokamorao ba sele [5085]	cell background color ...
mmala wa bokamorao ba leqephe [5084]	page background color ...
mmala o motshetla [5083]	grayscale ...
mmala o kganyang [5082]	bright colour ...
mmala o hlakileng [5081]	plain colour ...
mmala o fifetseng [5080]	dark colour ...
mmala o etetsweng wa kopantshano [5079]	visited link color ...
mmala [1] [5078]	coloured ...
mmala (s.3) mebala (pl.4) [2] [5077]	colour ...
Malawi [4763]	Malawi ...
malaria [4762]	malaria ...
mala [4761]	entrails ...
Lemalei (s.5) Mamalei (pl.6) [4163]	Malay person ...
Lemalei (s.5) Mamalei (pl.6) [4162]	Malay person ...
lemala [4161]	injured ...

Figure 7.3.2: Extraction of mala 'entrails' from Bukantswe Online (http://bukantswe.sesotho.org/display.php?action=search&word=mala&type=full)

The quality of treatment of *mala* stands in sharp contrast with other good electronic dictionaries. Compare, for example the treatment of 'entrails' in the Macmillan Dictionary (figure 7.3.3)



Did you know?

Click any word in a definition or example to find the entry for that word

entrails - definition and synonyms

NOUN [PLURAL] /ˈentreɪlz/

the organs inside the body of a person or animal, especially the intestines

Synonyms and related words

General regions and organs inside the body:
apparatus, canal, cavity...

This is the British English definition of **entrails**. View American English definition of **entrails**.
Change your default dictionary to American English.
View the pronunciation for **entrails**.

From the Blog

What is binge-watching?

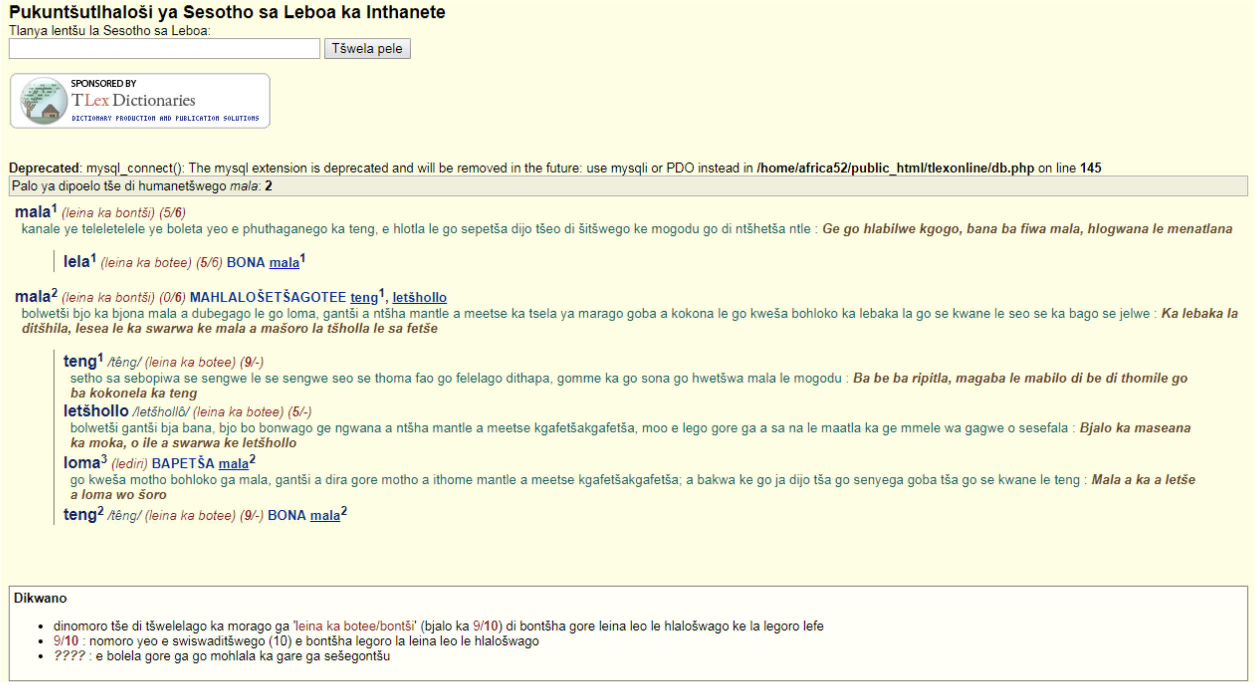
Is it 'suffragette' or 'suffragist'?

'Black and white' is a surprisingly colourful expres...

Figure 7.3.3: Extraction of entrails in the Macmillan Dictionary (https://www.macmillandictionary.com/dictionary/british/entrails)

The treatment includes part of speech information, a clickable pronunciation icon, phonetic transcription, a paraphrase of meaning, synonyms and related words, etc. This is a plausible treatment which Sesotho online dictionaries should follow because it does not carry just the meaning of 'entrails' but is surrounded by other important information with regards to the lemma. The user is presented with more information than they anticipated and in this way, they will learn from the dictionary. Furthermore, true electronic features (cf. Prinsloo 2003) such as pop-up boxes, audible pronunciation, links to the definitions of all words used in the paraphrase of meaning, thesaurus, etc., are utilised.

Consider also the treatment of *mala* in *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete* (https://africanlanguages.com/psl/)



Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete

Tlanya lentšu la Sesotho sa Leboa:

Deprecated: mysql_connect(): The mysql extension is deprecated and will be removed in the future: use mysqli or PDO instead in **/home/africa52/public_html/tlexonline/db.php** on line **145**

Palo ya dipoelo tše di humanetšwego *mala*: **2**

mala^{1} *(leina ka bontši) (5/6)*
kanale ye teleletelele ye boleta yeo e phuthaganego ka teng, e hlotla le go sepetša dijo tšeo di šitšwego ke mogodu go di ntšhetša ntle : ***Ge go hlabilwe kgogo, bana ba fiwa mala, hlogwana le menatlana***

lela^{1} *(leina ka botee) (5/6)* BONA mala^{1}

mala^{2} *(leina ka bontši) (0/6)* MAHLALOŠETŠAGOTEE teng^{1}, letšhollo
bolwetši bjo ka bjona mala a dubegago le go loma, gantši a ntšha mantle a meetse ka tsela ya marago goba a kokona le go kweša bohloko ka lebaka la go se kwane le seo se ka bago se jelwe : ***Ka lebaka la ditšhila, lesea le ka swarwa ke mala a mašoro la tšholla le sa fetše***

teng^{1} */têng/ (leina ka botee) (9/-)*
setho sa sebopiwa se sengwe le se sengwe seo se thoma fao go felelago dithapa, gomme ka go sona go hwetšwa mala le mogodu : ***Ba be ba ripitla, magaba le mabilo di be di thomile go ba kokonela ka teng***

letšhollo */letšhollô/ (leina ka botee) (5/-)*
bolwetši gantši bja bana, bjo bo bonwago ge ngwana a ntšha mantle a meetse kgafetšakgafetša, moo e lego gore ga a sa na le maatla ka ge mmele wa gagwe o sesefala : ***Bjalo ka maseana ka moka, o ile a swarwa ke letšhollo***

loma^{3} *(lediri)* BAPETŠA mala^{2}
go kweša motho bohloko ga mala, gantši a dira gore motho a ithome mantle a meetse kgafetšakgafetša; a bakwa ke go ja dijo tša go senyega goba tša go se kwane le teng : ***Mala a ka a letše a loma wo šoro***

teng^{2} */têng/ (leina ka botee) (9/-)* BONA mala^{2}

Dikwano

- dinomoro tše di tšwelelago ka morago ga 'leina ka botee/bontši' (bjalo ka 9/**10**) di bontšha gore leina leo le hlalošwago ke la legoro lefe
- 9/**10** : nomoro yeo e swiswaditšwego (10) e bontšha legoro la leina leo le hlalošwago
- *????* : e bolela gore ga go mohlala ka gare ga sešegontšu

Figure 7.3.4: Treatment of *mala* in Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete (https://africanlanguages.com/psl/)

This online dictionary has two (2) hits for *mala* with its plural and singular forms. Examples of usage have been clearly illustrated in bold in the different hits and paraphrased in such a way that the user is able to understand the true meaning of *mala*. Note also the elaborate guidance on the numbering system, which is lacking in the *Bukantswe Online* dictionary as noted above. This dictionary also has a user guide. In the text box with heading "*Dikwano*" the compilers explain what the numbers on the words mean.

Thus, a typical model entry of high lexicographic quality for a Sesotho electronic dictionary should entail true electronic feature, be elaborate, given the fact that there are no strict limitations on space in an electronic dictionary and utilise true electronic features.

mala

mala^1 (ma-la) bongata, Sehlopha sa 5/**6**

Tlh. Setho sa mmele se fumanwang ka mpeng ya motho kapa ka mpeng ya phoofolo se jarang dijo, ho di tsamaisa le ho di kenya maro a itseng a tsoang mabopong a ona.

Mohl. 1) Dikgoho di na le mala a masesane.
2) Batho bohle bana le mala
Bonngwe *lela* Sehlopha sa **5**/6

mala^2

Tlh. Lefu le tshwarang motho ha a dubehile ka mpeng, a jele dijo tse senyehileng kapa dijo di sa dula hantle ka mpeng.

Mohl. 1) Ntate o tshwerwe ke mala.
2) Mala a mmangwane a bohloko, ebile o a tsholla.
3) Nthabiseng o sebedisitse mala.

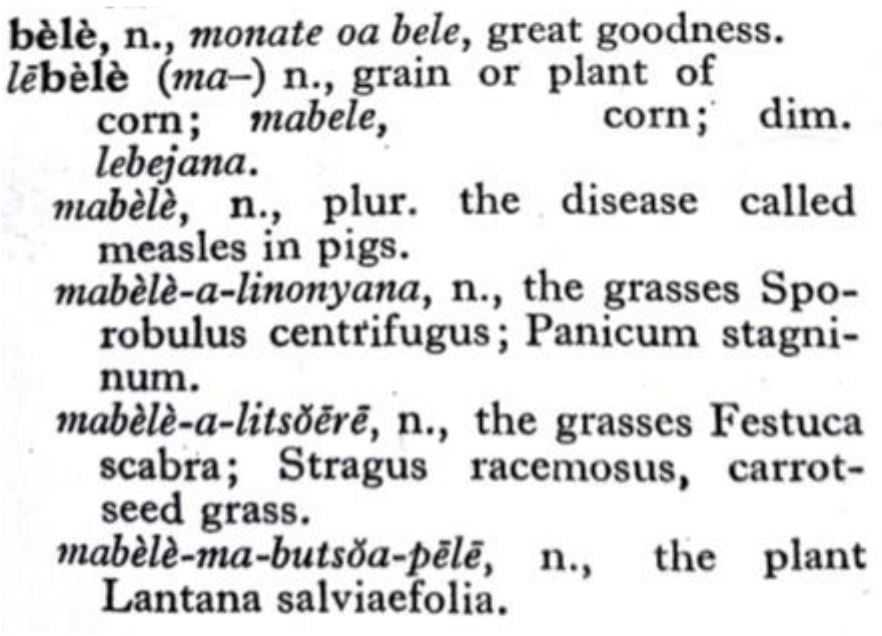
mala^3 Lela la pene

Mohl. 1) Pene ya Tshepo e na lela le le lelele
Mantswe a amanang le mala (mohodu, lela, lelana, letshollo, sokelwa)

By having dictionary articles arranged this way, or in a way similar to this, where the different definitions, examples of use, singularity and plurality indicated as well as related words are illustrated, Sesotho lexicography will be heading in the right direction.

7.4 Southern Sotho-English Dictionary

The Southern Sotho-English Dictionary is a bilingual dictionary published in 1950. This dictionary is a true reflection of dictionaries that were not compiled through the use of corpora. It uses Lesotho Sesotho orthography, which can be a problem for South African Sesotho learners who are possibly still learning the language. However, as a result of a lack of good quality, user-friendly dictionaries for Sesotho, users might opt to use it, with the hope of extracting good information from it.



bèlè, n., *monate oa bele*, great goodness.
*lē***bèlè** (*ma–*) n., grain or plant of corn; *mabele*, corn; dim. *lebejana*.
mabèlè, n., plur. the disease called measles in pigs.
mabèlè-a-linonyana, n., the grasses Sporobulus centrifugus; Panicum stagninum.
mabèlè-a-litsŏĕrē, n., the grasses Festuca scabra; Stragus racemosus, carrot-seed grass.
mabèlè-ma-butsŏa-pēlē, n., the plant Lantana salviaefolia.

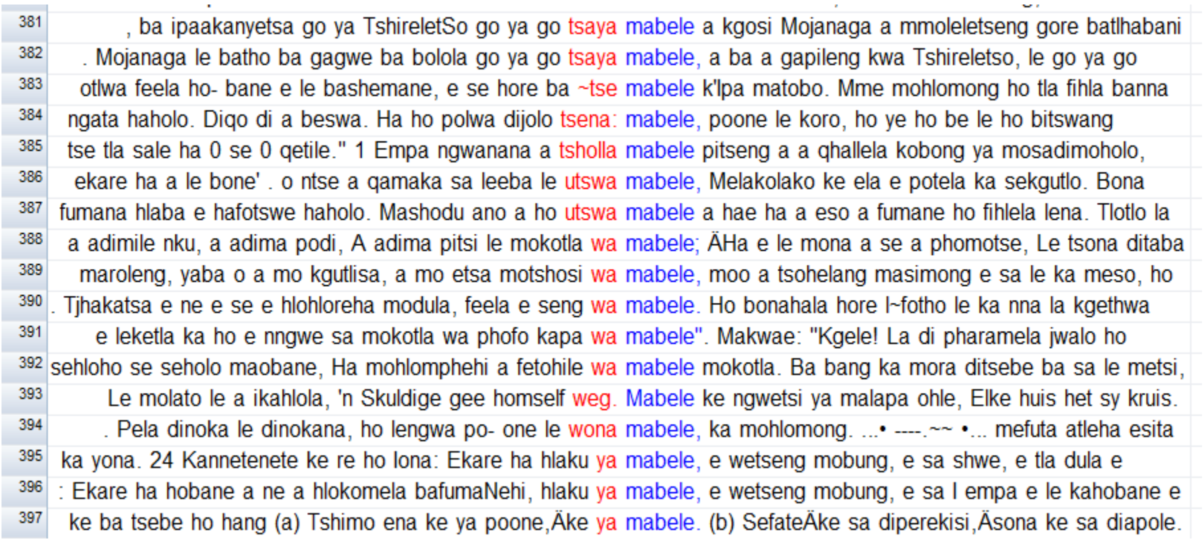
Figure 7.4.1: Extraction of bele and lebele from Mabille and Dieterlen (1950) Bilingual Dictionary (English -Sesotho) (offensive word in the description of *corn* has been deleted – see original source for full translation).

The entry in this extract is *bele,* which is indicated as a noun with the primary meaning of something that is good. Hence the example *monate oa bele* 'great goodness'. The prefix of the singular noun class 5 'le' is added to *bele* in italics to denote the word *lebele* 'grain or plant of corn' according to Mabille and Dieterlen **(**1950**).** The prefix of noun class 6 'ma-' is put in brackets. The user is alerted about the plurality of *lebele* through the prefix 'ma-'. However, they need to have previous knowledge about class prefixes and the noun class system and its numbering, or else they will be left confused. Under *lebele, mabele* is also treated and its diminutive form illustrated which is *lebejana*. Other meanings of *mabele* are shown, however, what is interesting is the way *mabele* is treated in the other definitions. In the second definition the compilers stated its part of speech but used the abbreviation for plural which is *plur.* Quite interestingly they didn't show singularity of *lebele* in the beginning. They assumed that the user already knows the different noun classes. The other definitions are meanings of certain plants and a disease respectively, however, examples of usage are not indicated in any of the definitions provided. This is problematic especially for learners of Sesotho. They need to know how words are used and in what context, but this dictionary is not giving them that. There is no etymology, pronunciation, synonyms of words, etc. Had a corpus been used in compiling this dictionary, these inconsistencies would have been eliminated. Consider, for example the richness of information on *lebele/mabele* given in the Sesotho corpus in figure 7.4.2 concordance lines. Also consider the following examples that have been extracted from figure 7.4.2. They are good examples of usage that can be used in the dictionary, that will allow for proper understanding of the word *mabele* and how it is used in context.

Pela dinoka le dinokana, ho lengwa poone le wona ***mabele*** *ka mohlomong mefuta atleha esita*

E leketla ka ho e nngwe sa mokotla wa phofo kapa wa ***mabele,*** *Makwae: Kgele! La di pharamela jwalo ho*

A adimile nku, a adima podi, a adima pitsi le mokotla wa ***mabele****; Aha e le mona a se a phomotse, le tsona ditaba*



381	, ba ipaakanyetsa go ya TshireletSo go ya go tsaya mabele a kgosi Mojanaga a mmoleletseng gore batlhabani
382	. Mojanaga le batho ba gagwe ba bolola go ya go tsaya mabele, a ba a gapileng kwa Tshireletso, le go ya go
383	otlwa feela ho- bane e le bashemane, e se hore ba ~tse mabele k'lpa matobo. Mme mohlomong ho tla fihla banna
384	ngata haholo. Diqo di a beswa. Ha ho polwa dijolo tsena: mabele, poone le koro, ho ye ho be le ho bitswang
385	tse tla sale ha 0 se 0 qetile." 1 Empa ngwanana a tsholla mabele pitseng a a qhallela kobong ya mosadimoholo,
386	ekare ha a le bone' . o ntse a qamaka sa leeba le utswa mabele, Melakolako ke ela e potela ka sekgutlo. Bona
387	fumana hlaba e hafotswe haholo. Mashodu ano a ho utswa mabele a hae ha a eso a fumane ho fihlela lena. Tlotlo la
388	a adimile nku, a adima podi, A adima pitsi le mokotla wa mabele; ÄHa e le mona a se a phomotse, Le tsona ditaba
389	maroleng, yaba o a mo kgutlisa, a mo etsa motshosi wa mabele, moo a tsohelang masimong e sa le ka meso, ho
390	. Tjhakatsa e ne e se e hlohloreha modula, feela e seng wa mabele. Ho bonahala hore l~fotho le ka nna la kgethwa
391	e leketla ka ho e nngwe sa mokotla wa phofo kapa wa mabele". Makwae: "Kgele! La di pharamela jwalo ho
392	sehloho se seholo maobane, Ha mohlomphehi a fetohile wa mabele mokotla. Ba bang ka mora ditsebe ba sa le metsi,
393	Le molato le a ikahlola, 'n Skuldige gee homself weg. Mabele ke ngwetsi ya malapa ohle, Elke huis het sy kruis.
394	. Pela dinoka le dinokana, ho lengwa po- one le wona mabele, ka mohlomong. ...• ----.~~ •... mefuta atleha esita
395	ka yona. 24 Kannetenete ke re ho lona: Ekare ha hlaku ya mabele, e wetseng mobung, e sa shwe, e tla dula e
396	: Ekare ha hobane a ne a hlokomela bafumaNehi, hlaku ya mabele, e wetseng mobung, e sa l empa e le kahobane e
397	ke ba tsebe ho hang (a) Tshimo ena ke ya poone,Äke ya mabele. (b) SefateÄke sa diperekisi,Äsona ke sa diapole.

Figure 7.4.2: Concordance lines for *lebele/mabele*

By having concordance lines such as these, the lexicographer is able to extract examples of use, frequency of use, different senses, collocations, etc. Also look at concordance lines for *kgasa* 'crawl' on figure 7.4.3, and together with figure 7.4.2, it is evident that such data can only be obtained through the use of a corpus. A large corpus will extract more information and a lot of concordance lines. Hence, Prinsloo (2015:293) notes that:

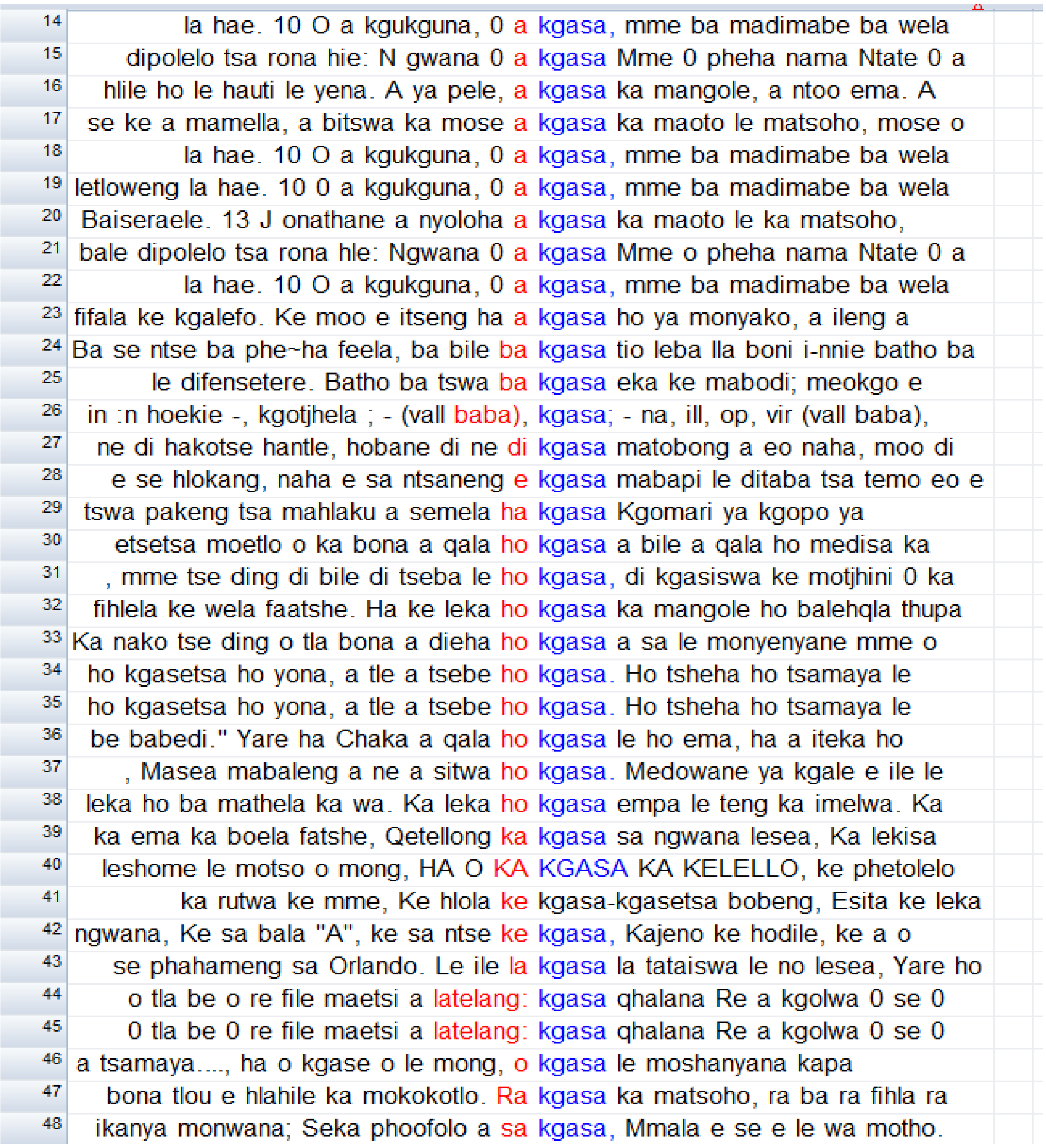
> …it is not feasible for the lexicographer to read through thousands of concordance lines generated for a specific keyword in context –100-300 lines could be regarded as a reasonable number to consider for detecting senses and to find typical collocations and authentic examples of use.

Consider the following concordance lines for *kgasa* extracted from, figure 7.4.3. These concordance lines reveal homonyms and senses for the word *kgasa* which can be used in the dictionary as examples. See the following the examples. *Kgasa* on line 16 has a different meaning to the one on line 40. The former means to 'crawl' while the latter referrers low thinking capacity.

14. La hae. 10 O a kgukguna, o a ***kgasa****, mme ba madimabe ba wela*

16. hlile ho le hauti le yena. A ya pele, a ***kgasa*** *ka mangole, a ntoo ema. A*

40. leshome le motso o mong, HA O KA ***KGASA*** *KA KELELLO, ke phetolelo*



14	la hae. 10 O a kgukguna, 0 a kgasa, mme ba madimabe ba wela
15	dipolelo tsa rona hie: N gwana 0 a kgasa Mme 0 pheha nama Ntate 0 a
16	hlile ho le hauti le yena. A ya pele, a kgasa ka mangole, a ntoo ema. A
17	se ke a mamella, a bitswa ka mose a kgasa ka maoto le matsoho, mose o
18	la hae. 10 O a kgukguna, 0 a kgasa, mme ba madimabe ba wela
19	letloweng la hae. 10 0 a kgukguna, 0 a kgasa, mme ba madimabe ba wela
20	Baiseraele. 13 J onathane a nyoloha a kgasa ka maoto le ka matsoho,
21	bale dipolelo tsa rona hle: Ngwana 0 a kgasa Mme o pheha nama Ntate 0 a
22	la hae. 10 O a kgukguna, 0 a kgasa, mme ba madimabe ba wela
23	fifala ke kgalefo. Ke moo e itseng ha a kgasa ho ya monyako, a ileng a
24	Ba se ntse ba phe~ha feela, ba bile ba kgasa tio leba lla boni i-nnie batho ba
25	le difensetere. Batho ba tswa ba kgasa eka ke mabodi; meokgo e
26	in :n hoekie -, kgotjhela ; - (vall baba), kgasa; - na, ill, op, vir (vall baba),
27	ne di hakotse hantle, hobane di ne di kgasa matobong a eo naha, moo di
28	e se hlokang, naha e sa ntsaneng e kgasa mabapi le ditaba tsa temo eo e
29	tswa pakeng tsa mahlaku a semela ha kgasa Kgomari ya kgopo ya
30	etsetsa moetlo o ka bona a qala ho kgasa a bile a qala ho medisa ka
31	, mme tse ding di bile di tseba le ho kgasa, di kgasiswa ke motjhini 0 ka
32	fihlela ke wela faatshe. Ha ke leka ho kgasa ka mangole ho balehqla thupa
33	Ka nako tse ding o tla bona a dieha ho kgasa a sa le monyenyane mme o
34	ho kgasetsa ho yona, a tle a tsebe ho kgasa. Ho tsheha ho tsamaya le
35	ho kgasetsa ho yona, a tle a tsebe ho kgasa. Ho tsheha ho tsamaya le
36	be babedi." Yare ha Chaka a qala ho kgasa le ho ema, ha a iteka ho
37	, Masea mabaleng a ne a sitwa ho kgasa. Medowane ya kgale e ile le
38	leka ho ba mathela ka wa. Ka leka ho kgasa empa le teng ka imelwa. Ka
39	ka ema ka boela fatshe, Qetellong ka kgasa sa ngwana lesea, Ka lekisa
40	leshome le motso o mong, HA O KA KGASA KA KELELLO, ke phetolelo
41	ka rutwa ke mme, Ke hlola ke kgasa-kgasetsa bobeng, Esita ke leka
42	ngwana, Ke sa bala "A", ke sa ntse ke kgasa, Kajeno ke hodile, ke a o
43	se phahameng sa Orlando. Le ile la kgasa la tataiswa le no lesea, Yare ho
44	o tla be o re file maetsi a latelang: kgasa qhalana Re a kgolwa 0 se 0
45	0 tla be 0 re file maetsi a latelang: kgasa qhalana Re a kgolwa 0 se 0
46	a tsamaya...., ha o kgase o le mong, o kgasa le moshanyana kapa
47	bona tlou e hlahile ka mokokotlo. Ra kgasa ka matsoho, ra ba ra fihla ra
48	ikanya monwana; Seka phoofolo a sa kgasa, Mmala e se e le wa motho.

Figure 7.4.3: Concordance lines for *kgasa* ‘crawl’

7.5 Cross-referencing

> The mediostructure, that is the system of cross-referencing, is a lexicographic device that can be used to establish relations between different components of a dictionary. Gouws and Prinsloo (1998:18)

In terms of Gouws and Prinsloo (2005) the user is referred from a reference position in the dictionary mostly by means of a reference marker to a reference address where (s)he can find more information. Consider the following example: Comparing to this example given by Gouws and Prinsloo (1998:20) from *Collins Dictionary of the English Language*.

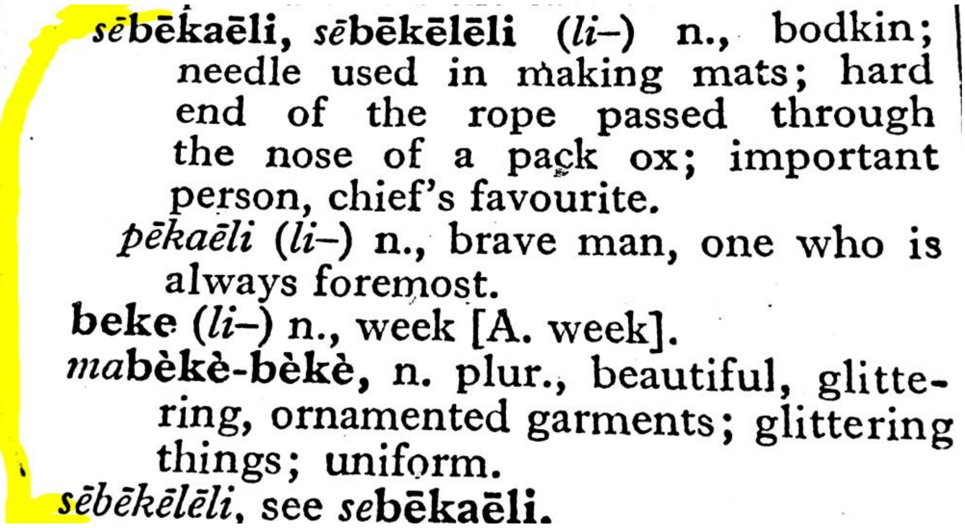
> gy·ro ('d3amm) *n., pI.* ·ros. 1. See gyrocompass. 2. See gyroscope

> The lexical item *gyro* is polysemous and has two different senses. The article of this lemma sign displays no meaning paraphrase for either of the polysemous senses but cross-refers the user instead to the treatment presented for two other lemma signs i.e. *gyrocompass* and *gyroscope.*

The example given by Gouws and Prinsloo (1998:20) above takes the user to the cross-reference of *gyro* because the lemma *gyro* is polysemous and as a result the user will find the meaning either on *gyrocompass* or *gyroscope.* The scenario is different for *sebekaeli* and *sebekeleli* in figure 7.5.1 because meaning was already inferred for both lemmas in the first treatment of *Sebekaeli.* It is a case where inappropriate cross-referencing has been done. The two lemmas at the beginning and end of the yellow highlight (figure 7.5.1) indicate a cross-reference according to Mabille and Dieterlen (1950).

However, the lemmas illustrated there are no different. The top lemmas are the same as the bottom ones. If there was a different lemma at the bottom that had the same definition as the one on top, that would make a difference and qualify to be a cross reference. This cross-referencing does not carry any meaning because there is no difference on what is written on the first line and what is written on the last. In this case, space was wasted which could have been used for other words most likely to be looked for. These kinds of discrepancies are avoided when proper lexicographical planning is adhered to, because that is where important decisions like what order should be followed and how to lemmatise lemmas takes place. Hence Gouws and Prinsloo (1998:24) believe: that:

> Cross-referencing has not been employed to its full potential in dictionaries for most African languages.



*sē***bēkaēli**, *sē***bēkēlēli** (*li–*) **n.**, bodkin; needle used in making mats; hard end of the rope passed through the nose of a pack ox; important person, chief's favourite.

pēkaēli (*li–*) **n.**, brave man, one who is always foremost.

beke (*li–*) n., week [A. week].

*ma***bèkè-bèkè**, n. plur., beautiful, glittering, ornamented garments; glittering things; uniform.

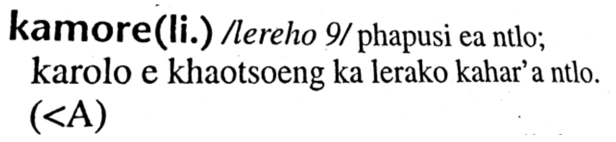
sēbēkēlēli, see *se***bēkaēli.**

Figure 7.5:1 Extraction of *sebekaeli* from English -Sesotho Bilingual Dictionary by Mabille and Dieterlen (1950)

Secondly, the order of the lemmas is also put in an ad hoc manner because the top lemma starts with *se*- so how did the dictionary team come to lemmatise *beke* ‘week’ and *mabeke-beke* ‘beautiful, glitter’ then go back to *se*- which was already lemmatised in the beginning?

7.6 Sethantšo sa Sesotho

This is a monolingual dictionary written in Lesotho Sesotho orthography. It was first published in 2005, when corpora were already being used in big English dictionaries. However, for Sesotho dictionaries and other African language dictionaries in general, the situation was dire and in need of urgent rescue and resuscitation. In an ideal world, the dictionary article presented in figure 7.6.1 was supposed to contain more information for the user. However, it only tells the user that *kamore* ‘room’ is a noun of class 9 and its plural form is indicated by (li.). No examples of usage, no etymology, no phonological descriptions or help with pronunciation, etc. “As such, African-language scholars … need to mirror the great contemporary endeavours in corpus linguistics achieved by scholars of, say, Indo-European languages for them to have the same impact that big dictionaries that used corpora have.” Prinsloo and de Schryver (2001:130)

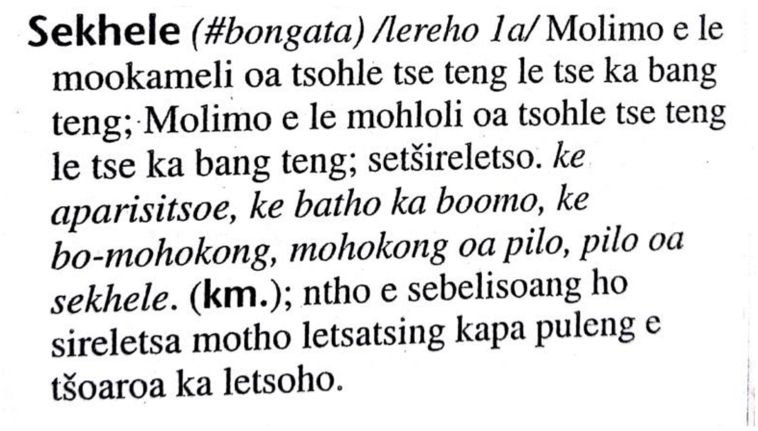


kamore(li.) */lereho 9/* phapusi ea ntlo;
karolo e khaotsoeng ka lerako kahar'a ntlo.
(<A)

Figure 7.6.1: Extraction of *kamore* ‘room’ from Sethantšo sa Sesotho Monolingual Dictionary (Sesotho) by Batho Hlalele (2005)

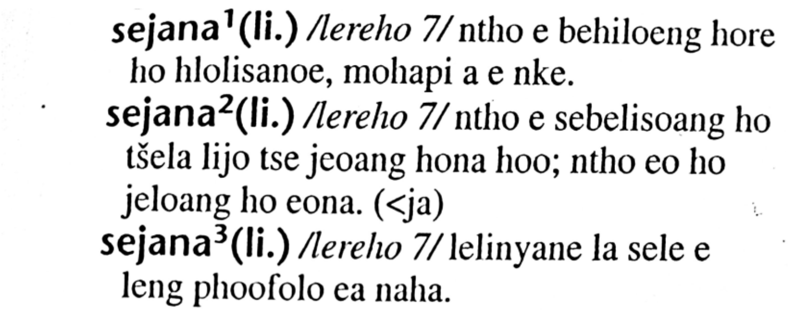
Collocations could have been used, to find words that normally go together with *kamore* ‘room’ to give users a clearer picture of what is meant by this lemma. That definition could have been accompanied by other secondary definitions like ‘bedroom’ to give the user other possible definitions and even given cross-references to *phaposi* ‘room’.

Consider the following example (figure 7.6.2). The first definition of *Sekhele* refers to God. Being a mother tongue speaker of Sesotho myself, I am doubtful of this definition, especially put as a primary definition. This instance already stipulates that there is a difference in my understanding as the researcher and the understanding of the compiler of the dictionary. Maybe some speech communities refer to God as *Sekhele* but this already shows that if corpora were used, these differences in understanding would not be present because the compiler would have gotten some kind of language representation and then lemmatised those words which had a high frequency standing. The second meaning is not represented as such, of ‘umbrella’.



Sekhele *(#bongata) /lereho 1a/* Molimo e le mookameli oa tsohle tse teng le tse ka bang teng; Molimo e le mohloli oa tsohle tse teng le tse ka bang teng; setšireletso. *ke aparisitsoe, ke batho ka boomo, ke bo-mohokong, mohokong oa pilo, pilo oa sekhele.* (**km.**); ntho e sebelisoang ho sireletsa motho letsatsing kapa puleng e tšoaroa ka letsoho.

Figure 7.6.2: Extract of *sekhele* ‘umbrella’ from Sethantšo sa Sesotho Monolingual Dictionary (Sesotho) by Batho Hlalele (2005)



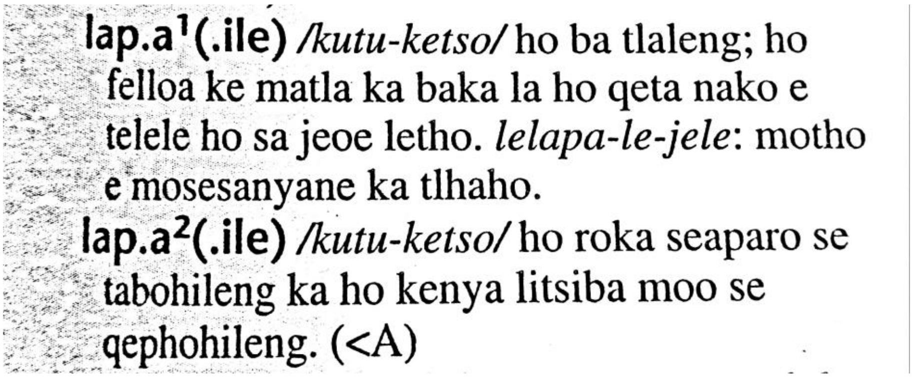
sejana^1(li.) */lereho 7/* ntho e behiloeng hore ho hlolisanoe, mohapi a e nke.

sejana^2(li.) */lereho 7/* ntho e sebelisoang ho tšela lijo tse jeoang hona hoo; ntho eo ho jeloang ho eona. (<ja)

sejana^3(li.) */lereho 7/* lelinyane la sele e leng phoofolo ea naha.

Figure 7.6.3: Extract of *sejana* from Sethantšo sa Sesotho Monolingual Dictionary (Sesotho) by Batho Hlalele (2005)

Compare figure 7.6.2 with 7.6.3 with regards to the way meanings are treated in the same dictionary. There are three different meanings of the lemma *sejana* 1) cup, as used in world cup 2) dish, 3) a type of wild animal. In figure 7.6.3, the definitions are put in order from primary to the other secondary meanings. The parts of speech and class have been indicated for each instance, whereas in figure 7.6.2 no part of speech has been assigned to either definitions.



lap.a^{1}(.ile) */kutu-ketso/* ho ba tlaleng; ho felloa ke matla ka baka la ho qeta nako e telele ho sa jeoe letho. *lelapa-le-jele*: motho e mosesanyane ka tlhaho.

lap.a^{2}(.ile) */kutu-ketso/* ho roka seaparo se tabohileng ka ho kenya litsiba moo se qephohileng. (<A)

Figure 7.6.4: Extract of *lapa* from Sethantšo sa Sesotho Monolingual Dictionary (Sesotho) by Batho Hlalele (2005)

Another example to look at is how the verb *lapa* 'hungry' has been treated in this dictionary. On the primary definition, the compilers indicated the suffix of the lemma in past tense, (-*ile*) then shows its part of speech *ketso* 'verb'. The first definition according to them *ho ba tlaleng* is not a true definition. That could actually be used as a typical example of usage and not a definition. More could have been added to that meaning if a corpus was used, especially the issue on examples of use. The same goes for the second definition, because it too does not have any examples of usage.

A brief comparison of pages 101-103 of Sethantšo sa Sesotho with a Sesotho corpus immediately reveal the inadequacies of lemma selection. Highly used Sesotho words (frequencies given in brackets) such as *latswa* (54) 'taste', *laya* (67) 'advice', *lebenkele* (45) 'shop', *lebese* (131) 'milk', *leeto* (117) 'journey', *letsopa* 'clay' (82), *letshwao* 'mark' (72) *letshehadi* 'left' (92), *letswai* 'salt' (75), *letswele* 'breast' (52), *lewatleng* 'at the ocean'(49) etc., should be properly treated because they are most likely to be looked for by users..

7.7 Conclusion

It can be concluded that the inconsistencies raised are a true reflection of African language dictionaries and show lack of proper lexicographic planning. Even though lexicographic planning is an important aspect of dictionary compilation, "when one considers the themes of papers delivered at national and international lexicographic conferences and congresses over the past decade, it is apparent that the planning and management of lexicographic projects is a theme which does not often arise". Van Schalkwyk (1999: 198)

Gouws and Prinsloo (2005: 9) also state that:

> The publication of any dictionary should not only be the result of the preceding compilation activities but it has to be regarded as the culmination of a much more comprehensive set of activities, the so-called lexicographic process. The compilation and eventual publication of any dictionary form part of at least one lexicographic process.

Furthermore, the inconsistencies addressed in this chapter reflect the need for Sesotho lexicographers to use corpora in dictionary compilation in order for them to take their rightful place in modern lexicographic trends and raise the use and status of their language. The three (3) dictionaries, Sethantšo sa Sesotho, Southern Sotho-English Dictionary and Bukantswe Online have been critically analysed and the results show that more work needs to be done in Sesotho dictionaries, not just to improve the language but to also make sure that they qualify for use in formal settings and users trust in what is presented to them. This will also help shape and save Sesotho. South African indigenous languages may be saved from the list of those identified by the UN Educational Scientific and Cultural Organisation (UNESCO) as likely to disappear at the end of this century. (Langa, 2017)

Chapter 8: Summary, conclusion and recommendations

8.1 Summary

The dictionary-making process in the pre-corpus era was complicated and was prone to errors, inaccuracies and gaps. Lexicographers used to rely mainly on their intuition and their inadequate, if not primitive, methodologies of manually consulting sources. Methods like reading, marking, etc., were used for many centuries, and this meant that lexicographers were the main source of information in the dictionary-compilation process. However, humans have their limitations – therefore, relying mainly on lexicographers' knowledge and intuition for a process as complex as compiling a dictionary is risky and exposes the process to problems concerning omission and inclusion of words. This meant that words most likely to be looked for by users were omitted and those not popular to them would also not be lemmatised. Simply put, if a lexicographer did not know a particular word or had not come across it, that word was not lemmatised. Other factors that play a role in the omission of some key words were related to physical constraints like fatigue, because the methods that were used at the time were laborious. This dissertation has proved exhaustively that it is humanly impossible for one individual to know and master all the words of a fully developed language. Hence, there is a need for the utilization of corpora and corpus query tools.

The corpus era brought new possibilities, exciting efficiencies and advantages, which led to lexicographers getting help with just a click of a mouse or touch of a screen. Corpora have the ability to draw on all instances of a word, including examples of use, frequency of use, study words in context, etc., and this in turn helps lexicographers enhance the quality of newly compiled or revised dictionaries. All these ensure that words most likely to be looked for by users are lemmatised. The idea is for Sesotho lexicographers to use this important tool, which has revolutionized dictionary compilation and language studies in general.

Sesotho is one of the official languages of South Africa that were previously marginalized and, together with other indigenous languages, it gained official status in 1994 when the country became democratic. It was one of the first African languages to be reduced to writing by the French missionaries, but it is disappointing that it is still not a fully functional language which can be used, with confidence, for academic, scientific, technological, medical, legal and other related advancement purposes. What is even more disappointing is the fact that Sesotho lexicographers and/or specialists have not produced, or made efforts to compile, good comprehensive dictionaries that are user-friendly.

8.2 Aims and conclusions

The aims of this dissertation were:

1. to provide a perspective on Sesotho in terms of its position as an African language and its place within the Bantu-language family in particular,
2. to provide an overview of the core aspects of dictionary compilation,
3. to make a critical evaluation of African-language dictionaries, with specific focus on Sesotho dictionaries, and
4. to provide recommendations and attempt to lay the foundation for the compilation of corpus-based dictionaries of high lexicographic quality for Sesotho, focusing on macrostructural and microstructural aspects.

These aims have been addressed in the different chapters in the dissertation.

In chapter 2, the focus was on Sesotho as an African language of the Bantu-language family in South Africa. In this chapter it was explained that it is part of the Niger-Congo languages which fall under the South-eastern Bantu zone. It falls under the Sesotho cluster, which includes Setswana, Sepedi and Sesotho. It was one of the first languages to be reduced to writing by the missionaries and is spoken mostly in Lesotho and South Africa (with the Free State having the highest number of Sesotho speakers). Nonetheless, the orthography of Sesotho in Lesotho and in South Africa varies a lot to the extent that scholars from these two countries are unable to share scholar and academic work produced in Sesotho. This language has different dialects ranging from Sekwena, Sephuthi, Setlokwa, Setaung Sekgolokwe and Serotse, and a few clicks which are mostly derived from the ‘q’ sound, as in the case of ‘qh’ and other sounds already mentioned. Sesotho is promoted by *Lesedi FM* and *Bona* magazine in South Africa.

In chapter 3, the history of corpora was discussed, together with the advantages of using this modern lexicographic tool that changed the study of language – and more precisely the study of lexicography. It has been suggested that the utilization of corpora and corpora tools are an important aspect in compiling dictionaries of a high lexicographic standard, hence it is vital that Sesotho lexicographers take advantage of this great initiative. This chapter also highlighted the advantages, for lexicographers, of using corpora. Gone are the days when lexicographers had to rely on their intuition.

Corpora have revolutionised the compilation of macrostructure and microstructure of dictionaries and this was discussed in detail in chapters 4 and 5 respectively. In chapter 4, discussions on the structure of the macrostructure were dealt with and it was demonstrated that the richness (or lack of information) on the macrostructure affects user understanding. Corpora help lexicographers to separate the seldom-used words from the often-used words and also helps with compilation of lemma lists using frequency tests. Most importantly, in African languages, lexicographers have to negotiate a complex interplay between lemmatisation approaches, strategies, traditions, and nominal and verbal structures, together with degrees of conjunctivism versus disjunctivism.

In chapter 5, microstructural aspects were discussed in detail. This concerned the entry and the type of information that an entry should contain. Entry information like definitions, examples of usage, etymology, parts of speech, illustrations in some dictionaries, etc., were discussed – which all have the aim of making the entry understandable and useful to users.

Chapter 6 dealt with English Online dictionaries (Macmillan and Collins dictionary). These dictionaries were discussed because they are compiled with the use of corpora and are a benchmark against which Sesotho dictionaries are measured. They represent a methodology that should be followed in Sesotho dictionaries, whether online or printed. They do not only provide users with useful information, they are also excellent reference materials for teaching and learning, hence we call them perfect guides that can be used for Sesotho dictionaries. The evaluation of Sesotho dictionaries in Chapter 7 revealed many inconsistencies and inadequacies which could have been avoided had corpora been used. This chapter also revealed the lack of lexicographic planning and tenacity within Sesotho dictionaries which often leaves users confused, either because words they are looking for in the dictionary are not lemmatised, or because there are certain rules they have to know first before attempting to use the dictionary, or because in some instances dictionaries are presented without any user guide.

The South African Constitution of 1996 ensured the promotion and development of all South African languages, and this dissertation responds to that challenge, as it intends to add to the limited research in Sesotho-language studies and more specifically in Sesotho lexicography. This attempt is also a response to an invitation, or rather a challenge, set by De Schryver and Prinsloo (2000:304) when they posit that:

> The field of electronic corpus lexicography is not an esoteric branch of linguistics that can only be pursued by scholars of the English language. Even if the linguistic field of 'electronic corpus lexicography' was somehow invented by lexicographers of the English language, and even if it has been booming ever since for (British) English, corpus-aided and corpus-based dictionaries for the African languages are a reality too.

Much more work still needs to be executed in Sesotho lexicography to ensure that Sesotho is on par with other languages of the world, and this dissertation is only the beginning.

8.3 Recommendations

Utilisation of corpora is the map for the road ahead for Sesotho lexicography. Corpora play an important role in lemma selection and the compilation of lemma lists, hence Sesotho lexicographers cannot continue to ignore its importance. An example – such as the brief comparison of pages 101-103 of Sethantšo sa Sesotho dictionary with a Sesotho corpus – immediately reveal the inadequacies of lemma selection. Highly used Sesotho words (frequencies given in brackets) such as *latswa* (54) ‘taste’, *laya* (67) ‘advice’, *lebenkele* (45) ‘shop’, *lebese* (131) ‘milk’, *leeto* (117) ‘journey’, *letsopa* ‘clay’ (82), *letshwao* ‘mark’ (72) *letshehadi* ‘left’ (92), letswai ‘salt’ (75), *letswele* ‘breast’ (52), *lewatleng* ‘at the ocean’(49), etc., should be properly treated in Sesotho dictionaries because they are most likely to be looked for by users.

Going forward, other aspects of dictionary compilation that need to be explored in detail are the following:-

1. Compilation of front and back matters.

2. Coverage, layout and user-friendliness of the user’s guide.
3. Quality, consistency and variety in defining strategies.
4. Use of standard dictionary conventions.
5. Choice, ordering and composition of translation-equivalent paradigms.

These aspects form the core of dictionary compilation and exist mutually for the benefit of the user.

The challenges of compiling corpus-based dictionaries for Sesotho is attainable only if Sesotho scholars start realizing the role and importance of electronic corpora in language studies. If well-developed languages of the world were able to achieve this, it is also possible for Sesotho. The future is corpus-based, and corpus ruled. Let’s take our rightful place in the new millennium. Let’s improve, preserve and protect our language.

REFERENCES

ABBYY. (n.d.). Retrieved from https://www.abbyy.com/en-eu/finereader/ Accessed on 24 August 2018

Accredited Language Services. (n.d.). *Sesotho*. Retrieved from http://www.alsintl.com/resources/languages/Sesotho/ Accessed on 17 July 2018

Afrikan Heritage. (n.d.). *Map of Africa*. Retrieved from www.afrikanheritage.com/official-spoken-languages-of-african-countries/ Accessed on 05 February 2016

Adas, M. (2001). *Agricultural and Pastoral Societies in Ancient and Classical History*. Temple University Press.

Musanhu, B (2009) "Chapter 5: Lesotho" In Denis Kadima and Susan Booysen (eds) Compendium of Elections in Southern Africa 1989-2009: 20 Years of Multiparty Democracy, EISA, Johannesburg, 161-162. Retrieved from https://www.eisa.org.za/wep/lesmedia.htm Accessed on 10 October 2018

Al-Jarf, R (1998). *Dictionary Skills Course material*. Retrieval from https://www.researchgate.net/profile/Reima_Al-Jarf/publication/314094550_Dictionary_Skills/links/58b544a7a6fdcc6b2b31e1a2/Dictionary-Skills.pdf [25 March 2018]

Ally, S. and Lissoni, A. (Eds.). (2017). *New Histories of South Africa's Apartheid-Era Bantustans*. Routledge

Alonso Ramos, M., M. Garcìa Salido and O. Vincze. (2014). *Towards a Collocation Writing Assis-tant for Learners of Spanish*. Faaß, G. and J. Ruppenhofer (Eds.). 2014. Workshop Proceedings of the 12th Edition of the KONVENS Conference, Hildesheim, Germany, October 8–10, 2014: 77-88. Hildesheim: Universitätsverlag Hildesheim.

Atkins, B.T. Sue, Michael Rundell and Edmund Weiner. 1997. *Salex'97 Compilation of Monolingual Dictionaries*. Unpublished course material of a tutorial held at the Dictionary Unit for South African English, Rhodes University, Grahamstown, 15–26 September 1997.

Atkins, S., Clear, J. and Ostler, N. (1992). In McEnery, T. Xiao, R. Tono, Y.(Eds.). (2006). *Corpus-based Language Studies: An Advanced Resource Book*. Taylor & Francis.

Atkins, S. Clear, J. and Ostler, N (1991) Corpus Design Criteria. Retrieved from http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf Accessed on 2 July 2018

Austin, P. (2008). One Thousand Languages: Living, Endangered, and Lost. University of California Press.

Awak, M.K. (1990). *Historical Background, with Special Reference to Western Africa*. In R.R.K. Hartmann (Ed.). (1990) Lexicography in Africa: Progress Reports from the Dictionary Research Centre Workshop at Exeter. University of Exeter Press.

Bandyopadhyay, S. (1999). *Detection and correction of phonetic errors with a new orthographic dictionary*. Computer Science and Engineering Department Jadavpur University, Calcutta, INDIA.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Germany Gunter Narr Verlag

Bergenholtz, H. and Tarp, S (1995). *Manual of Specialised Lexicography: The preparation of specialised dictionaries*. John Benjamins. Amsterdam.

Berry, J. and Sebeok T. A. (2017). *Linguistics in Sub-Saharan Africa: Aus: Current Trends in Linguistics*. Walter de Gruyter GmbH & Co KG.

Biber, D. (1993). *'Representativeness in corpus design'. Literary and linguistic Computing 8/4*: 243-57 https://doi.org/10.1093/llc/8.4.243

Bradley, J., Bradley, L., Fine, V., Jon Vidar (2001). *South Africa: Lesotho & Swaziland*. Modern Overland.

Carroll, K. (ed.). (2012). *Collins COBUILD Advanced Dictionary of English*. Sewende uitgawe. Glasgow: HarperCollins.

Caxton Magazines. (n.d.). *BONA*. Retrieved from https://www.caxtonmags.co.za/bona Accessed on 12 February 2018

CCURL. (2014): Proceedings overview: *Workshop on collaboration and computing for under-resourced languages in the Linked Open Data Era*. Retrieved from http://www.ilc.cnr.it/ccurl2014/Charles Hamm (2006). *Putting Popular Music in Its Place*. Cambridge University Press.

Childs, T, G (2003). *An Introduction to African Languages*. John Benjamins Amsterdam.

Christian, L. (2017). *The structure of a dictionary*. Retrieved from https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/dict_structure.html Accessed on 25 June 2018

Collins free online Dictionary. (n.d.). *Translator*. Retrieved from https://www.collinsdictionary.com/translator Accessed on 14 May 2018

Collins. (n.d.) History of COBUILD. Retrieved from https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild) Accessed on 10 August 2018

Collins. (n.d.) English-French dictionary. Retrieved from (https://www.collinsdictionary.com/dictionary/english-french) Accessed on 14 May 2018

Collins, R. O. and Burns, J.M. (2007). *A History of Sub-Saharan Africa*. Cambridge University Press.

Constitution of the Republic of South Africa. (n.d.). Section 29. Education. Retrieved from http://www.usig.org/countryinfo/laws/South%20Africa/The%20Constitution%20of%20South%20Africa%2029.pdf Accessed on 05 May 2018

Constitution of the Republic of South Africa. South African Government (1996). Retrieved from http://www.elections.org.za/content/Documents/Laws-and-regulations/Constitution/Constitution-of-the-Republic-of-South-Africa,-1996/ 7 July 2018

Coulmas, F. (2013). *Sociolinguistics: The Study of Speakers' Choices*. Cambridge University Press. https://doi.org/10.1017/CBO9781139794732

Cowie A. P. (1998). *Phraseology: Theory, Analysis, and Applications*. OUP Oxford.

Crystal, D. (2014). *Language Death*. Cambridge University Press.

Davies, M. (n.d.). *Overview, search types, looking at variation, corpus-based resources*. https://corpus.byu.edu/size.asp Accessed on 10 August 2018

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Springer.

De Schryver, G.M and Prinsloo, D. (2000). *Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure*. 291–309 South African Journal of African Languages.20 (4). https://doi.org/10.1080/02572117.2000.10587437

De Schryver, G.M and Prinsloo, D. (2000). *Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure*. 310-330. South African Journal of African Languages. 20(4). https://doi.org/10.1080/02572117.2000.10587438

De Schryver G.-M. and Prinsloo, D.J. (2001). *Corpus-based Activities versus Intuition-based Compilations by Lexicographers. The Sepedi Lemma-sign List as a Case in Point*. Nordic Journal of African Studies. Volume 10. Number 3, pp. 374-398

De Schryver G.M (2003). *Online Dictionaries on the Internet: An Overview for the African Languages.* Lexikos, 13, 1-20.

De Schryver, G.M. and Prinsloo, D.J. (2003). *Compiling a lemma-sign list for a specific target user group: The Junior Dictionary as a case in point Dictionaries*. Journal of the Dictionary Society of North America *24*. 28-58

De Schryver, G. M (2007). *Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane*. E gatišitšwe ke Oxford. Cape Town: Oxford University Press Southern Africa.

De Schryver, G.M. (2007). *Oxford Bilingual School Dictionary: Northern Sotho*. Oxford University Press.

G.M. de Schryver (Ed). (2010) *Oxford Bilingual School Dictionary: Zulu and English*. First Edition. lviii + 582 pp. Cape Town: Oxford University Press Southern Africa.

Doke, C.M. and Mofokeng, S.M. (1957). Textbook of Southern Sotho Gramma. Cape Town: Longman. Retrieved from http://www.sesotho.web.za/sound.htm Accessed on 28 August 2018

Deuter, M., Bradbery J. and Turnbull, J. (Eds). (2015). *Oxford Advanced Learner's Dictionary of Current English*. 9th Edition. Oxford: Oxford University Press.

Dicts.info (n.d.) *Free dictionary projects.* Retrieved from https://www.dicts.info/dictionary.php?l1=English&l2=Sesotho&word=medicine&Search=Search Accessed on 04 October 2018

Free State Government. (2003). Retrieved on http://www.fs.gov.za/departments/sac/library/sesotho_creative_literature_jul-sept2003_profiling_literature.htm Accessed on 25 May 2018.

Fihlani, P. (2017. August 30). *Trying to save South Africa's first language*. BBC News. Retrieved from https://www.bbc.com/news/world-africa-39935150 Accessed on 02 August 2018

Francis, W. N. and H. Kučera. (1964). *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Frawley, W., Hill, K.C., Munro, P. (2002). *Making Dictionaries: Preserving Indigenous Languages of the Americas*. University of California Press.

Fuertes-Olivera, P., A. (2010). *Specialised Dictionaries for Learners*. Walter de Gruyter.

Gough. D.H. (n.d.). English In South Africa. Retrieved from http://www.salanguages.com/english/esa.htm Accessed on 20 June 2018

Gouws, R. H. (1990). *Information Categories in Dictionaries with Special Reference to Southern Africa*. In Hartmann, R.R.K. (Ed.). *Lexicography in Africa*. Exeter: University of Exeter Press, 52–65.

Gouws, R. H. (1996). *Bilingual Dictionaries and Communicative Equivalence for a Multilingual Society*. Lexikos *6*, 14-31

Gouws, R. H. and Prinsloo, D. J. (1998). *Cross-referencing as a lexicographic Device*. *Lexikos 8*. 17-36

Gouws, R. H. and Prinsloo, D. J. (2005). *Principles and Practice of South African Lexicography*. SUN PReSS

Gouws, R. Heid, U. Schweickard, W. Wiegand, H, E. (2013). *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on De Schryver, G.M., (2013)., *Electronic and Computational Lexicography. XIX. Computer-based Dictionary making II: Tools and Procedures*. Walter de Gruyter.

Granger, S and Paquot, M (2012). *Electronic Lexicography*. Oxford University Press.

Guthrie, M. (1971). *The western Bantu Languages*. In Sebeok, A.T., (Ed). *Current Trends in Linguistics*. Vol.7.357-366

Hartmann, R.R.K. and James, G (1998). *Dictionary of Lexicography*. Routledge

Hlalele, B. (2005) *Sethantšo sa Sesotho*. Longman. Maseru.

Hüllen, W. (2006). *English Dictionaries, 800-1700: The Topical Tradition*. Clarendon Press.

Jeffery C. (2000). *Projected Corpora of South Africa's Official Languages*. In De Schryver, G.M. and Prinsloo, D.J. (2000). *The compilation of electronic corpora, with special reference to the African Languages*. Southern African Linguistics and Applied Language Studies 2000 *18*: 89-106

Johanson, S and Hofland K. (1989). Frequency analysis of English vocabulary and grammar, based on the LOB Corpus, I: Tag frequencies and word frequencies. Oxford: Clarendon Press in Hudson, R (1994). About 37% of word-tokens are nouns. Vol. 70, No. 2, pp. 331-339. Linguistic Society of America. Retrieved from https://www.jstor.org/stable/pdf/415831.pdf?casa_token=oEe_hwA1ucUAAAAA:FaXgTzc6LB4UxqBVrWlolGzUfDss_qEJxr1-l6Z1ys1uuZtD-JOEDcWVmA1s4jIloqvAJK-ubCfm7ITz-kGsOfdyQb2iA0p_Kmaur2RBPjdx1XFyM3Dov Accessed on 04 April 2018

Jiang, W. (2000). *The Relationship Between Culture and Language*. ELT Journal. Vol.54/4. Oxford University Press.

Kennedy G. (1998). *An Introduction to Corpus Linguistic*s. London: Longman. In De Schryver, G.M. and Prinsloo, D.J. (2000). *The compilation of electronic corpora, with special reference to the African languages*. Southern African Linguistics and Applied Language Studies 2000 *18*: 89-106

Khumalo, L. (2015). *Advances in developing Corpora in African Languages*. Kuwala Acalan Journal • 1(2): 21-29.

Khumalo, L. (2017). *The design and implementation of a corpus management system for the isizulu national corpus*. In *African association for lexicography 22nd international conference* 26-29 June 2017. In cooperation with the conference of the language associations of southern Africa (CLASA) Rhodes university, Grahamstown, South Africa

Kilgarriff, A. (1997). Putting Frequencies in the Dictionary. International Journal of Lexicography 10(2): 135-155.

Lan, L (2005). *The growing prosperity of on-line dictionaries. English Today* 83, Vol. 21, No. 3. Cambridge University Press

Langa, P. S. L (2017-10-15). *New dictionaries to help preserve SA's language heritage*. City Press. Retrieved from https://www.news24.com/SouthAfrica/News/new-dictionaries-to-help-preserve-sas-language-heritage-20171014 Accessed on 14 August 2018

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.

Laviosa, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications.* Rodopi. Amsterdam

Leech, G.N (1992). *100 Million Words of English: The British National Corpus (BNC).** Retrieved from http://s-space.snu.ac.kr/bitstream /10371/85926/3/1.%202235197.pdf Accessed on 06 September 2018

Lesedi FM (n.d.). About Lesedi FM. Retrieved from www.lesedifm.co.za/sabc/home/lesedifm/aboutus Accessed on 13 August 2018

Lewis, M (2000). *Teaching Collocation: Further Developments in the Lexical Approach*. Vol.4. No.4. England: Language Teaching Publications. Retrieved from http://www.tesl-ej.org/wordpress/issues/volume4/ej16/ej16r12/?wscr Accessed on 21 October 2018

Lingling, L. and Hai, X (2015). *Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners*. Lexikos 26, 391-401.

Mabille and Dieterlen (1950). *Southern Sotho-English dictionary*. Morija. Basutoland: Morija Sesuto Book Depot. Lesotho

Macmillan Dictionary (n.d.). Retrieved from https://www.macmillandictionary.com Accessed on 16 April 2018

Macmillan dictionary(n.d.). `Entrails'. Retrieved from https://www.macmillandictionary.com/dictionary/british/entrails Accessed on 05 July 2018

Macmillan education (n.d.). STOP THE PRESS: Dictionary no longer a page-turner Retrieved from http://www.macmillaneducation.com/MediaArticle.aspx?id=1778 Accessed on 16 April 2018

May, K. and Ortlepp, A. (2009/2007). *Dictionaries, Concept and Use*. Retrieved from https://www.tuchemnitz.de/phil/english/sections/linguist/independent/kursmaterialien/dict/ortlepp.pdf Accessed on 20 August 2018

Mayor, M. (ed.). (2009). *Longman Dictionary of Contemporary English*. Vyfde uitgawe. Harlow, Essex: Pearson Education.

McEnery, T., & Xiao, R. (2010). W*hat corpora can offer in language teaching and learning*. In E. Hinkel (Ed.)., *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 364-380). London & New York: Routledge.

McEnery, T. Xiao, R. Tono, Y.(Eds.). (2006). *Corpus-based Language Studies: An Advanced Resource Book*. Taylor & Francis.

McIntosh, C. (ed.). (2013). *Cambridge Advanced Learner's Dictionary*. Vierde uitgawe. Cambridge: Cambridge University Press.

McLeod, S. (2010). *Laying the foundations for Multilingual acquisition: An International overview of speech acquisition*. In Cruz-Ferreira, M. (2010). *Multilingual Norms*. Peter Lang. New York

Mojela, M.V. (2007) (ed.). *Pukuntšutlhaloši ya Sesotho sa Leboa*. Pietermaritzburg: Nutrend.

Ngobeni, S (2010). *Scholarly Publishing in Africa: Opportunities & Impediments*. African Books Collective

Nielsen, S (1994). *The Bilingual LSP Dictionary: Principles and Practice for Legal Language*. Gunter Narr Verlag.

Nuance. (n.d.). *Omnipage*. Retrieved from, https://www.nuance.com/print-capture-and-pdf-solutions/optical-character-recognition/omnipage.html Accessed on 28 August 2018

Nugues, P. M. (2006). *An Introduction to Language Processing with Perl and Prolog: An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German*. Springer Science & Business Media

O'Keeffe, A. and McCarthy, M. (2010). *'Historical perspective: What are corpora and how have they evolved?'*, in O'Keeffe, A. & McCarthy, M. (Eds)., *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3-13.

Omniglot. (n.d.). *Southern Sotho (seSotho).* Retrieved from https://www.omniglot.com /writing/sesotho.htm Accessed on 25 July 2018

One Planet Nations Online. (n.d.). *Map of the Distribution of African Language Families and some Major African Languages*. Retrieved from www.nationsonline.org/ oneworld/map/african-langauge-map.htm Accessed on 15 January 2017

Otlogetswe, T.J. (2013a). *Introducing Tlhalosi ya Medi ya Setswana: The Design and Compilation of a Monolingual Setswana Dictionary*. Lexikos 23, 532-547.

Otlogetswe, T.J. (2013b). *Oxford English Setswana Setswana English School Dictionary*. OUP. Oxford.

Ortlepp, Anika (2007). Bilingual Dictionaries: General Facts and Macrostructure. Retrieved fromhttps://www.tuchemnitz.de/phil/english/sections/linguist/independent/ kursmaterialien/dict/ortlepp.pdf Accessed on 14 July 2018

Oxford Dictionary (n.d.). Retrieved from https://en.oxforddictionaries.com Accessed on 30 August 2018

Pasfield-Neofitou, S. (2009). *Paper, Electronic or Online? Different Dictionaries for Different Activities*. Vol.34, No.2. Babel. In Toyoda, E. (2016). *Usage and efficacy of electronic dictionaries for a language without word boundaries*. EUROCALL Review, Volume 24, No. 2 Retrieved from https://files.eric.ed.gov/fulltext/EJ1148659.pdf Accessed on 28 September 2018

Prinsloo, D. J (2013). *Issues in compiling dictionaries for African Languages*. In: Jackson, H (Ed.). (2013). *The Bloomsbury Companion to Lexicography*. A&C Black. London

Prinsloo, D. J, and De Schryver G, M. (1999). *The lemmatization of nouns in African languages with special reference to Sepedi and Cilubà*, SA Journal of African Languages, 19(4). https://doi.org/10.1080/02572117.1999.10587405

Prinsloo, D. J. (2001). *The Compilation of Electronic Dictionaries for the African Languages*. Lexikos 11:139-159

Prinsloo, D. J. and De Schryver G. M. (2000). *Taking Dictionaries for Bantu Languages into the New Millennium-with special reference to Kiswahili, Sepedi and isiZulu*. In Mdee, J.S. & H.J.M. Mwansoko (Eds.). 2001. Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings: 188–215. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.

Prinsloo, D. J., Bothma, T.J.D. and Heid U. (2013). *User support in e-dictionaries for complex grammatical structures in the Bantu languages*. Proceedings of the 16th Euralex International Congress. 15-19 July 2014. Bolzano. Italy

Prinsloo, D.J. (2005). *Electronic Dictionaries viewed from South Africa*. Hermes, Journal of Linguistics No 34.

Prinsloo, D.J. (2011). *A Critical Analysis of the Lemmatization of Nouns and Verbs in isiZulu*. Lexikos 21:169-193.

Prinsloo, D.J. (2015). *Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus*. Lexikos 25, 285-300.

Prinsloo, D.J. (2016). *A Critical Analysis of Multilingual Dictionaries*. Lexikos 26: 220-240.

Prinsloo, D.J. (2012). *Electronic lexicography for lesser-resourced languages: The South African context*. eLexicography. Sylviane Granger & Magali Paquot (Eds.). Oxford: Oxford University Press. 119-144.

Prinsloo, D.J. (2015) *Analysing words as a social enterprise: Lexicography in Africa with specific reference to South Africa*. Australex 2015 19-21 November 2015. Massey University, Albany Campus Auckland, Aotearoa/New Zealand.

Prinsloo, D.J., Bothma, T.J.D. & Heid, U. (2013). *Linking e-dictionaries to advanced processed corpus data*. eLex 2013, Tallinn, 17-19.

Prinsloo, D.J., Bothma, T.J.D., Heid, U. and 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. (403-426).

Prinsloo, Danie J. and Bosch, Sonja E. (2012). *Kinship terminology in English–Zulu/Northern Sotho dictionaries - a challenge for the Bantu lexicographer*. In Ruth Vatvedt Fjeld & Julie Matilde Torjusen (Eds.). Proceedings of the 15th Euralex International Congress. 7-11 August 2012. Oslo. 296-303

Ramadiro, B. (2013). *African languages in the print media*. Retrieved from https://www.litnet.co.za/african-languages-in-the-print-media/ Accessed on 30 June 2018

Readiris. (n.d.). Retrieved from (http://www.irislink.com/EN-US/c1729/Readiris-17--the-PDF-and-OCR-solution-for-Windows-.aspx). Accessed on 28 August 2018

Renouf A. (1987). Corpus Development. In Sinclair JM (Ed). *Looking Up. An account of the Cobuild Project in lexical computing and the development of the Collins Cobuild English Language Dictionary*. London: Collins ELT. 1-40.

Research Centre Workshop at Exeter, 24–26 March 1989: Exeter Linguistic Studies 15 (pp. 8–18). Exeter, England: University of Exeter Press.

Rialland, A., Riddouane, R., Hulst H. (2015). *Features in Phonology and Phonetics: Posthumous Writing*s by Nick Clements and Co-authors Walter de Gruyter GmbH & Co KG. France

Robert Allen. (2008). *Lumping and splitting*. Cambridge University Press.

Rundell, M. (2015). *From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions*. Lexikos 25, 301-322.

Rundell, M. (Ed.). (2007). *Macmillan English Dictionary for Advanced Learners*. Tweede uitgawe. Oxford: Macmillan.

Rundell. M (2013, March 12). *Macmillan Dictionary: Our move from print to online - your questions answered*. Retrieved from https://www.youtube.com/watch?v=PGK4QTDsA_Q Accessed on 14 July 2018

Scott, M. (1996). *Wordsmith Tools: Manual*. Oxford: Oxford University Press.

Sebolela, F. (2009). *The compilation of corpus-based Setswana dictionaries*. Unpublished Doctoral thesis. University of Pretoria.

Selçuk University *Journal of Faculty of Letters*. 1300-4921 (Print).; 2458-908X (Online). Selçuk Üniversitesi Edebiyat Fakültesi Dergisi

Sesotho Online. (n.d.). *Nouns & Noun classes*. Retrieved from http://www.sesotho.web.za/nouns.htm Accessed on 21 April 2018

Sesotho Online. (n.d.) *Bukantswe v.3 - Online Sesotho - English dictionary -- Bukantswe ya Sesotho le Senyesemane*. Retrieved from http://bukantswe.sesotho.org/ Accessed on 21 April 2018

Sinclair, J. (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. Retrieved from https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm Accessed on 11 August 2018

Snyman D.S, J.W. Shole, J.S. and Le Roux, J.C. (1990). *Dikišinare ya Setswana English Afrikaans Dictionary*. Pretoria: Via Afrika.

Snyman, Jannie W. (ed.). (1990). *Dikišinare ya Setswana – English – Afrikaans Dictionary / Woordeboek*. Pretoria: Via Afrika Limited.

South African Centre for Digital Language Resources (SADiLaR). (n.d.). Retrieved from https://www.sadilar.org/ Accessed on 01 September 2018

Stefanowitsch, A (2003). *Corpus Design*. Retrieved from https://www.dbthueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00010791/corp_design.pdf Accessed on 10 March 2018

Summers, D. (1993). Longman Lancaster English Language Corpus Criteria and Design. International Journal of Lexicography, Vol.6, No.3.

The Nilo-Saharan Language Family. (n.d.). Retrieved from http://www.ling.fju.edu.tw/typology/Nilo-Saharan.htm Accessed on 12 March 2017

The United Nations Educational, Scientific and Cultural Organization (UNESCO). (n.d.). *UNESCO's Endangered Languages Programme*. Retrieved from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/FlyerEndangeredLanguages-WebVersion.pdf Accessed on 30 April 2017

TshwaneDJe Software and Consulting.(n.d.). *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete*. Retrieved from https://africanlanguages.com/psl/ Accessed on 26 August 2018

University of Free State. Faculty of the Humanities. African Languages. (n.d.). Sesiu sa Sesotho National Lexicography Unit Retrieved from https://www.ufs.ac.za/humanities/departments-and-divisions/african-languages-home/unlisted-pages/development-centre-and-sesiu-sa-sesotho-national/sesiu-sa-sesotho-national-lexicography-unit-2 Accessed on 01 June 2018Van Schalkwyk, D.J. (1999). Planning and Management - *the Most Neglected Activities in Lexicography*. Lexikos 9: 198-207.

Van Sterkenburg, P (2003). *A Practical Guide to Lexicography*. John Benjamins Amsterdam.

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

Verlinde, S., P. Leroyer, and J. Binon. (2010). Search and You Will Find. *From Stand-alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats*. International Journal of Lexicography *23*(1) 1-17

Vetulani, Z., Uszkoreit, H., and Kubis, M. (2016). Human Language Technology. *Challenges for Computer Science and Linguistics*: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers. Springer

Wadsworth Cengage Learning. (n.d.). *Annotated dictionary entry*. Retrieved from http://college.cengage.com/english/raimes/keys_writers/5e/assets/students/annotated_dictionary.html Accessed on 15 July 2018

Wang Dakun (2001). Should they look it up? *The role of dictionaries in language learning*. REACT. 1, 27-33. National Institute of Education (Singapore).

Wanner, L., S. Verlinde and M. Alonso Ramos. (2013). Writing Assistants and Automatic Lexical Error Correction: Word Combinatorics. Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia: 472-487. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Webster's Online Dictionary (2011). Retrieved from http://www.thecre.com/emerging/20110306.html Accessed on 15 April 2019

Ziervogel D, Mokgokong PC (1975). Comprehensive Northern Sotho Dictionary. Pretoria: J. L. van Schaik. (CALD). McIntosh, C. (Ed.). 2013. Cambridge Advanced Learner's Dictionary. Vierde uitgawe. Cambridge: Cambridge University Press.