University of the Witwatersrand

School of Literature, Language and Media

Towards the Development of a Monolingual Sesotho Learner's Dictionary

By

Malefu J. Mahloane

475894

A Masters dissertation submitted in fulfilment of the requirements for the degree of Master of Arts in African Languages & Linguistics

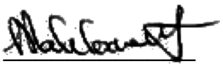
Supervisor: Ms Q.T.M. Setshedi

January 2023

Plagiarism Declaration

I, Malefu J. Mahloane, declare that this research is my original initiative. All sources consulted have been appropriately referenced. The work is submitted to the University of the Witwatersrand and has not been submitted to any other institution.

Date: 08/01/2023

Signature: 

Acknowledgements

My sincere gratitude goes to my Supervisor, Ms Q.T.M. Setshedi for seeing me through this process of establishing myself as an Academic, even after her retirement. For believing in me and pouring out her expertise to ensure that the work becomes what it is today – despite life's challenges.

Thanks to Prof. Tawana Kupe who arranged for the SLLM department to purchase a laptop for me in the early stages of this research.

I am eternally grateful to the Erasmus-Mundus Aesop programme, the University of the Witwatersrand, and the National Research Foundation for the financial support.

Thanks to my grandfather, Ntatemoholo Ramonne Mahloane who inspired the initiation of the dictionary project, and to Mr M.J. Ntisa, Mr M.A. Mabena and Dr N. Mdlalose for their unconditional support.

Lastly, this is for Ntatemoholo Tanki so Ntisa my late grandfather, best friend and cheerleader who inspired me to continue, no matter what.

Ke a leboha. Asante.

Abstract

Sesotho practical lexicography commenced over 160 years ago and up to the time when this research was initiated, there had not been a single monolingual or online learner's dictionary produced for the language. As a result, this research aimed to develop a Sesotho monolingual learner's dictionary with the ultimate goal being to make the dictionary available online. To achieve this, the research investigated the different factors to consider when developing a dictionary with these characteristics. The objective of this research was to use Corpus Linguistics methodology to develop the dictionary. The methods used are both qualitative and quantitative. The contents of the dictionary are extracted from a c.a. 100 000-word corpus that was compiled specially for this research – as a corpus is the object of study or a tool in Corpus Linguistics. The corpus was processed using various language processing methods and tools, the main being the *TshwaneLex* Suite. The study managed to develop a prototype of the Sesotho monolingual learner's dictionary and it is attached as Annexure 1, and the dictionary will continue to be developed beyond this research. It is anticipated that, the developed prototype dictionary presented here, will serve as evidence that there are effective means that Sesotho lexicographers can adapt to produce other types of dictionaries beyond the usual bilingual/translation and paperback dictionaries that are the norm in Sesotho lexicography.

Table of Contents

Plagiarism Declaration ii
Acknowledgements iii
Abstract iv
List of Figures viii
List of Tables ix
CHAPTER 1: INTRODUCTION 1
1.1 Background to Study 1
1.1.1. Lexicography in General 1
1.1.2. Sesotho Lexicography 7
1.1.3. Corpus Linguistics 14
1.2 Statement of the Problem 15
1.3 Aims, goals, rationale and research questions 19
1.3.1. Aim of research 19
1.3.2. Goal(s) of research 19
1.3.3. Rationale 19
1.3.4. Research Questions 20
1.4 Framework and Research Design 20
1.5 Results: their significance, practicability, and generalisability 23
1.6 Scope 24
1.7 Chapter Outline 25
CHAPTER 2: LITERATURE REVIEW 29
Introduction 29
2.1 Historical background of Sesotho 29
2.1.1. Overview of Sesotho 29
2.1.2. The Orthography of Sesotho 33
Table 2.1: *RSA and Lesotho orthography differences* 34
2.1.3. The semantic theory of Sesotho 35
2.2 The morphology of Sesotho 38
Table 2.2: *Sesotho Parts of Speech categories* 38
2.2.1. The Morphology of the noun 38
Table 2.3: *Noun classes* 39

Table 2.4: *Agreement/Concord markers of the noun and the qualificative in Sesotho* 40
2.2.2. Verb Structure 43
Table 2.5: *Types of verbal extensions* 44
2.2.3. The Qualificatives 44
Table 2.6: *Types of qualificatives* 45
2.2.4. Morphology of the Adverb 48
2.2.5. The Pronoun 49
2.2.6. Fixed Categories 51
2.2.7. Hybrid classes 52
2.3 Corpus Linguistics in lexicography 53
2.3.1. Lexicography 60
2.3.2. Bantu languages lexicography 65
2.3.3. Sesotho Lexicography 66
Table 2.7a: *Timeline and Survey table of Sesotho Dictionaries 1876-1965* 69
Table 2.7b: *Timeline and Survey table of Sesotho Dictionaries 1965-2018* 71
2.3.4. Dictionary making theories 71
2.4 Dictionary types and classifications 74
2.5 Chapter summary 78
CHAPTER 3: METHODOLOGIES AND DATA COLLECTION 80
Introduction 80
3.1 Methodologies 80
3.2 Corpus 85
3.2.1. Nature and size of compiled corpus 85
3.2.2. Machine readability 87
3.2.3. Sampling, balance and representativeness 87
3.3 Tools for data processing and dictionary design 92
3.3.1. Corpus compilation Tools 92
3.3.2. Dictionary making software: *TLex Suite* 94
3.3.3. Photocopier, scanner and camera 96
3.4 Methods/Steps 96
3.4.1. Corpus compilation 96
3.4.2. Steps for dictionary design 97
3.5 Scope of study 97
3.6 Summary of chapter 98
CHAPTER 4: DATA ANALYSIS AND RESULTS 99
Introduction 99

4.1 Corpus Analysis 99
Table 4.1: *Texts sampled for the corpus* 100
Table 4.2: *TEI-LITE Header* 103
4.2 Customizing *TLex* for *Sehlalosi Sa Sesotho sa Baithuti* 110
Table 4.2.1: *Sesotho POS names and their glosses* 113
4.3 Dictionary analysis 115
4.3.1. Macrostructure 115
4.4 Results 136
4.5 Conclusion 137
CHAPTER 5: DISCUSSION OF RESULTS AND CONCLUSION 138
Introduction 138
5.1. Interpretation of results 138
5.1.2. What kind of data and tools can be used to develop a monolingual online learner's dictionary? 145
5.1.3. Final research question 147
Table 5.1.3.1: *Comparison of Headwords/Lemmata (Macrostructure) with online dictionaries* 155
5.2 Results: Significance, practicality of study and recommendations. 155
5.3. Implications of Study 157
5.4 Limitations 159
5.5 Future Work 160
5.6. Conclusion of study 160
5.6.1. Brief statement of the problem 160
5.6.2. Aim and Research Questions 161
5.6.3. Rationale, Goal(s) 161
5.6.4. Methods used for analysis 162
REFERENCE LIST 164
Annexure 1: The Prototype Dictionary – Sehlalosi sa Sesotho sa Baithuti 177

List of Figures

Table of Contents v
Figure 2.1: *Niger-Congo family tree* 30
Figure 2.2: *South-eastern Bantu languages* 31
Figure 2.3: *Areas where Sesotho is Spoken* 32
Figure 2.4: *Online Sesotho - English dictionary – Bukantswe ya Sesotho le Senyesemane* 77
Figure 4.1: *Distribution of text genres in the corpus* 101
Figure 4.2: *Scanned text sample from Kgapha tsa ka (Maphalla: 1993)* 104
Figure 4.3: *Auto-converted version of the Poem in Figure 4.2.* 105
Figure 4.4: Edited version of 4.3. 106
Figure 4.5: *Sample of Mochochonono newspaper* 106
Figure 4.6: *Typed Article from Mochochonono* 107
Figure 4.7: Example of *code-mixing in Sesotho texts* 108
Figure 4.8: *Manually typed hymn sample* 109
Figure 4.2.1: *TLex interface for insetting dictionary properties* 111
Figure 4.2.2: *Replace English POS names with Sesotho ones* 112
Figure 4.2.3: *POS in Sesotho now as the default language* 112
Figure 4.2.4: *Names of cross-reference types changed to Sesotho* 114
Figure 4.3.1: *Macrostructure of Sehlalosi sa Baithuti* 116
Figure 4.3.2: *tlCorpus wordlist – high frequency words* 117
Figure 4.3.3: *Mid-frequency words* 118
Figure 4.3.4: *Low frequency words* 119
Figure 4.3.5: *POS distribution in the prototype dictionary i.e., Annexure 1* 119
Figure 4.3.6: TLex pop-up menu with lemma properties statistics, including total number of homonyms inSehlalosi 122
Figure 4.3.7: *TLex Interface for creating entries* 123
Figure 4.3.8: *tlCorpus concordance interface where usage examples were extracted* 123
4.3.9: *TLex entries interface – Identifying items with cross-references* 124
Figure 4.3.2.1: *Mark-up of nouns* 125
Figure 4.3.2.2: *Mark-up of reduplicated noun* 126
Figure 4.3.2.3: *Mark-up of derived nouns – Part 1* 127
Figure 4.3.2.4: *Mark-up of derived nouns – Part 2* 128
Figure 4.3.2.5: *Mark-up of a word with a cross-reference* 129
Figure 4.3.2.6: *Mark-up of agreement concord* 130
Figure 4.3.2.7: *Mark-up of lemmas belonging to more than one POS/function* 131
Figure 4.3.2.8: *Mark-up of qualificatives* 132
Figure 4.3.2.9: *Mark-up of a primitive verb and its associated derivative and word forms* 133
Figure 4.3.2.10: *Mark-up of derived verbs* 134
Figure 4.3.2.11: *Mark-up of different types of adverbs in the dictionary* 134

List of Tables

Table 2.1: ***RSA and Lesotho orthography differences*** **34**
Table 2.2: ***Sesotho Parts of Speech categories*** **38**
Table 2.3: ***Noun classes*** **39**
Table 2.4: ***Agreement/Concord markers of the noun and the qualificative in Sesotho*** **40**
Table 2.5: ***Types of verbal extensions*** **44**
Table 2.6: ***Types of qualificatives*** **45**
Table 2.7a: ***Timeline and Survey table of Sesotho Dictionaries 1876-1965*** **69**
Table 2.7b: ***Timeline and Survey table of Sesotho Dictionaries 1965-2018*** **71**
Table 4.1: ***Texts sampled for the corpus*** **100**
Table 4.2: ***TEI-LITE Header*** **103**
Table 4.2.1: ***Sesotho POS names and their glosses*** **113**
Table 5.1.3.1: ***Comparison of Headwords/Lemmata (Macrostructure) with online dictionaries*** **155**

CHAPTER 1: INTRODUCTION

This Chapter seeks to provide an overview of the study. The aim of the study was to develop a pilot monolingual online Sesotho Learner's dictionary using the corpus-based approach, based upon the long-standing shortage of monolingual and learner's dictionary in Sesotho. The methodologies adapted for the purpose of the study are Lexicography and Corpus Linguistics. In its basic sense, Lexicography is the making and studying of dictionaries. On the other hand, Corpus Linguistics is a methodology that is used to study language from corpora, it has two approaches, which are the corpus-driven and corpus-based approaches. To achieve the aim of the study, the objectives were set out to compile a Sesotho corpus, which is the main tool in the study, and utilise language processing tools that are recommended for the creation of an online dictionary. After exploring and assessing various such tools, the *TshwaneLex* Dictionary Suite was used for the creation of the dictionary. The prototype Sesotho Monolingual Online Learner's dictionary is presented alongside this report and further work is being done towards a larger version of the dictionary. The prototype is reflective of the macrostructure and microstructure of the envisaged product, which shall in the future be made available to the public.

1.1 Background to Study

This section will cover Corpus linguistics, Lexicography generally and Sesotho Lexicography in particular to provide a backdrop to the research.

1.1.1. Lexicography in General

Lexicography is as old as the beginnings of writing. Considine (2019) traces lexicography to Sumerian wordlists in Cuneiform, that is, clay tablets that are the precursors of writing as we know it today to circa 3200BC and it spread in the Mesopotamia area and its neighbours and finally to all over the world. A resolution on 22 March 1634 by the members of the newly founded

Académie françoise (subsequently Académie française) to devote themselves to the cultivation of the French language. By early 1635 they had agreed that this might be undertaken in part by the making of a dictionary, on which work had begun by 1637 and would continue until its publication in 1694. As such, lexicography is both ancient and modern because of the developments that have taken place in its development. Further, he highlights the fact that wordlists are the basis of lexicography whether ancient or modern. Notwithstanding, in this instalment the focus is not on the history of lexicography but on showing that it is a discipline with a very long history and has grown to become an independent discipline in its own right even though it is closely linked to other disciples. Attention is not going to be focused on the arguments about whether lexicography really is independent or not.

Closer to home, according to Durkin (2016), the *Dictionnaire de l'Académie* of French continued the initial work on the French Dictionary, which had started in 1637, and published the first French dictionary in 1694. English lexicography and practice, however, dates as far back as 1755 starting with the work of Samuel Jackson, who published two English dictionaries in that year (Osselton in Hartmann, 1983: 13). Osselton (in Hartmann 1983: 13) also points out that dictionaries first came into being in response to very practical needs such as recording the vocabulary of a language, an off shoot of writing. According to Rey (1995), lexicography practice started in the 1700s for European languages with the vision to describe language. Describing a language is important if the language is taught and learnt academically, thus making it pivotal to create resources that illustrate the use and structure of the language. An additional factor is the meaning creation of lexical items in the language to enable its users to acquire the words thoroughly and critically apply or express themselves in the language.

Consecutively, the pioneers carried out lexicographic projects for different languages as a dedication to their development. Rey (1995) states that the need for compiling dictionaries

became socially significant when the need arose to relate different languages to each other was defined, especially, during the 17th Century in Spain, Italy, France, and English as the self-regulatory description of a language. It may be asked at this stage whether there is a need to relate different languages, to which the answer is yes, seeing that there are developments in science and technology that are conceptualized, mainly, in major world languages making translations into other languages a requirement.

The simplest definition of lexicography is that it refers to the compilation of dictionaries, however, this definition is too general and does not do enough to clarify this concept which is much more complex. Bergenholtz and Gouws (2012:33) analysed five definitions of lexicography and found four of them unsatisfactory and so identified with the lengthy Wikipedia definition below which he felt was fairly balanced and covers the scope of the discipline even though it does not refer to dictionaries as practical tools:

> Lexicography is the study of lexicons, and is divided into two separate but equally important academic disciplines: **Practical lexicography** is the art or craft of compiling, writing and editing dictionaries. **Theoretical lexicography** is the scholarly discipline of analysing and describing the semantic, syntagmatic and paradigmatic relationships within the lexicon (vocabulary) of a language, developing theories of dictionary components and structures linking the data in dictionaries, the needs for information by users in specific types of situation, and how users may best access the data incorporated in printed and electronic dictionaries. This is sometimes referred to as 'metalexicography'. (http://en.wikipedia.org/wiki/Lexicography)

He simplified it here as - *Lexicography is divided into two related disciplines: practical lexicography, which is the art or craft of compiling of dictionaries, and also theoretical lexicography or metalexicography, which is the scholarly discipline of analysing and describing lexicons* - because it is the most accommodative in spite of its one weakness – that it does not refer to dictionaries as practical tools. This suggests that definitions are varied and though they may differ, there is general consensus about the nature of lexicography – it is a

coin with two sides that are interdependent: theoretical lexicography or general lexicography or metalexicography and practical lexicography.

This subject area is independent but can be applied across disciplines such as linguistics and information science amongst others. All the disciplines that support lexicography add to the wholesome implication of lexicography with the components that stipulate the processes, tools and methods involved in both the practice and theory of lexicography. Lexicography is closely linked to linguistics as shown by the definition of McEnery (1992:25) of lexicography as the study of occurring word tokens, falling under the morphological aspect of linguistics, as morphology allows for analysis of semantic and grammatical class membership. To corroborate this, Wierzbicka (1985) maintains that semantics has an obligation to provide theoretical foundations for lexicography. This is exact as dictionaries are information tools and semantics comprises aspects of sense, reference, implication and logical form. These aspects can be deduced from a single word token in a dictionary entry. Hartmann (1996) maintains that the criteria for an applied linguistic discipline are that "it should be linguistic in orientation, interdisciplinary in outlook and problem-solving in spirit" (in Hartmann 2001:33). Therefore, lexicography qualifies as a linguistic discipline. The preceding conflicts prove that there is divided agreement on the status of lexicography, and this is not surprising as the discipline was not clearly defined until the latter part of the 20th century as shall be shown lately.

Lexicography is also closely related to Information Science as both deal with data, information and knowledge and end-users who are human beings in both cases and similarities end there as the objectives differ. It can equally be placed under information science as it is in applied linguistics when considering that it is a method of preserving language, that indicates word classes and from which language use and structure can be studied in both print and electronic media.

Jackson (2002) adds that the practice of lexicography has always drawn on a range of other disciplines and crafts, relevant to reference works, it is widely considered as either a discipline in its own right (*sui generis*) or a branch of reference science or information science. Bergenholtz and Gouws (2012: 76) however, maintain that lexicography is a part of information science and not a sub-discipline of linguistics. They classify it in relation to information science rather than placing emphasis on the storage aspect of lexicographic information, that is, lexicography develops from the scientific study of language and social language needs. This results in the compilation of lexicons, additionally, as reference tools from which information can be retrieved. The definitions above may be different in terms of the stance taken and whether it is a discipline in its own right or not. However, as matters stand currently, lexicography is a discipline in its own right intersected by linguistics and information science and other disciples and subsequently, there is a body of literature underpinning it. Further, though it is inseparable from other disciplines, it has reached maturity and continues to be expanded.

The scope of lexicography is vast, especially, metalexicography as captured in the fifth definition in Bergenholtz and Gouws (2012: 33) quoted from Wikipedia as follows:

> Theoretical lexicography is the scholarly discipline of analyzing and describing the semantic, syntagmatic and paradigmatic relationships within the lexicon (vocabulary) of a language, developing theories of dictionary components and structures linking the data in dictionaries, the needs for information by users in specific types of situations, and how users may best access the data incorporated in printed and electronic dictionaries. This is sometimes referred to as 'metalexicography'. (http://en.wikipedia.org/wiki/Lexicography)

According to Gouws & Prinsloo (2005: 1), the two-sided nature of lexicography as known today became apparent in the latter part of the 20th Century. The theoretical component can be regarded ‘as a late-comer because lexicography was originally only associated with the practice

of dictionary making and as such up until the twentieth century, the practical component of lexicography was dominant' they observe. This scope arises from the scholars widely regarded as pioneers of modern lexicography in the latter part of the 20th Century as ancient lexicography had no clearly defined theory. These scholars are Zgusta and Wiegand respectively.

According to Imbaraywo (2008) the former, whose most influential work is dated 1971, is credited with having established theoretical lexicography as a research field and bringing into the fold of linguistics thus clarifying the link between itself and linguistics. Zgusta (1971:15) argues that "… a lexicographer needs to be familiar with linguistics in a much broader sense and has to take into consideration not only the whole structure of the language in question, but also the culture of the respective linguistic community." The latter does not only clearly establish the link between lexicography and linguistics (Gouws & Prinsloo 2005:2) but also reinforces lexicography as a worthy area of research. The decade following Zgusta's work, the 1980s to 1990s was dominated by Wiegand, who provided the general theory regarding the contents of dictionaries, dictionary articles and structure of dictionaries in the main and highlighted the fact that the object of lexicography is dictionaries and not language though the influence of linguistics permeates it. Wiegand (1983) further adds that the lexicographical practices and the results of these practices are components of focus in lexicography. The execution of lexicographic practices includes planning, establishing a dictionary base as well as processing the base in a lexicographical file. Dictionary compilation should involve consultation of other language resources. Imbayarwo (2008: 4) adds that the scope of lexicography also covers dictionaries and their public, history and theory of lexicography, problems of descriptions, dictionary types, dictionary criticism. These 'heuristic tools' are important in creating better dictionaries. Finally, the main purpose of the general or theoretical lexicography is to guide lexicographers to practical lexicography, the act of

compiling dictionaries (ibid. p.17). This scope attests to the varied nature of lexicography but also confirms its independent status.

Lexicography theory has grown in leaps and bounds in response to technological developments ushered in by the introduction of a personal computer in the 1960-1970s and so compiling dictionaries is now also electronic and has added online dictionaries to paper ones. Sesotho is behind in this regard as the biggest known bilingual, quick-reference online dictionary is still in its infant stage in terms of its macro and micro structures at about 10,007 entries and so still requires extensive work. Nonetheless, this is a step forward.

In short, in spite of varied views on whether lexicography is independent or applied, lexicography can be understood as a discipline of its own, in its own right; a discipline that is interdisciplinary, but not a strict information science or linguistic or other discipline.

1.1.2. Sesotho Lexicography

Though Motjope-Mokhali et al. (2020: 261) state that Sesotho lexicography began in the 1870s with the production of a bilingual English – Sesotho wordlist that was published in Morija, Lesotho, research on the beginnings of Sesotho lexicographic tradition by the biographers of the pioneers of Sesotho lexicography actually shows that it started as an individual's wordlist creation in 1852 on a very long ship journey from France to Lesotho via Cape Town by a Parish missionary, Rev Adolph Mabille in preparation for his primary assignment of evangelising the Basotho (Mabille and Dieterlen (1979: iii-vii) which was published twenty-six years later in 1876/78. As stated earlier, the development of lexicography for any given language is always related to the reduction of a language into writing. The first Parisian missionary arrived in Lesotho in 1831 and by 1841, Rev. Eugene Casalis had written a grammar book titled *Etudes sur la Langue Séchuana* but he had not added a list of words to this book. Though Rev Mabille's list was meant for his own language learning effort, it grew exponentially and by

1876/78 it was published as a *Sesuto-English Vocabulary* booklet consisting of 158 single column pages. By the time its 3rd edition was ready to be published in 1894, he died, and his work was continued by a co-worker, Rev Dieterlen who was familiar with his work and the dictionary finally reached the 8th edition with a total of 20039 entries in double format entries covering varied subject fields of the time in 1961. From the 4th edition onwards, the list was renamed *Sesuto-English Dictionary* and has retained this name to date and is generally considered the most reliable authority in Sesotho lexicography. As such, the list started by Mabille and expanded by Dieterlen with the help of community members, up and coming Basotho scholars and according to Kiango (2000:2) colonial administrators in South Africa is the bedrock of Sesotho lexicography. These dictionaries are bilingual and of a general-purpose nature for all users from basic level of education to professoriate and their sizes vary from 158 to 535 pages and so are voluminous. All editions were prepared and published in Morija except the 8th edition which was published at Thabana Morena, Lesotho. The development of this first Sesotho dictionary developed over a period of 109 years as a side work for missionaries purely for the development of Sesotho – "… their admiration for a great treasure and wish to preserve it" (Mabille and Dieterlen (1979: ix).

Throughout the development of these editions, Sesotho had only one orthography but in the late 1950s in preparation for the proclamation of South Africa as a republic, South Africa adopted a reformed orthography of Sesotho. Any Mosotho can read both orthographies without effort as the changes are minimal. See 2.1.2. The Orthography of Sesotho for details about the changes in the orthography and ever since the two orthographies have lived side by side. One of the impacts of this changed political status of South Africa was that the 7th edition was also published in the South African orthography in 1959. Further, these dictionaries were meant for communication between the missionaries and the Basotho and so are second language speaker oriented.

Further, Hair (1970) states the following with regard to the 1870s – 1990s African Languages dictionaries in general that:

> "Except for a few, all other vocabularies lack systematic arrangement. Some were collected to serve as conversation manual for European traders and sailors but are poorly organized; the majority were collected perhaps only as curiosities, funny words from a strange land." (in Hartmann, 1990: 10)

Their arrangement of the macrostructure is stem-based for three of the major lexical classes – nouns, verbs and qualificatives (that is, all types of adjectives) and this renders them difficult to use as they require sound morphological knowledge of Sesotho. The other weakness of dictionaries in this series is its alphabetical arrangement of vocabulary stretches which is phonemic, based on the initial sound instead of the first letter of the alphabet of the word. In spite of these few cons, it has been the model of all subsequent dictionaries. Though this reference is authoritative and was last revised in 1974, it has not been updated since the 8th edition in the late 1950s and with the advent of technological developments, language has grown immensely and this dictionary falls short with current vocabulary, especially, the technological one.

In addition to Mabille and Mabille and Dieterlen dictionaries of 1878-1974, Sesotho dictionaries have continued to be produced mainly in bilingual format even by native speakers of Sesotho such as Motsapi (2015) and the Sesotho Lexicographic Unit (2018) even though it is the 21st Century and there is a dearth of Sesotho dictionaries targeted at mother-tongue speakers. A survey of the history of all dictionaries of Sesotho done by Setaka and Prinsloo (2020:448) covers only the period from 1905 to 2017. The author researched the period before 1905 for fuller coverage (see 2.3.3. Sesotho Lexicography) for the list. The analysis of this survey follows in a latter paragraph below. Both surveys are inclusive of both print and online dictionaries. Both surveys attest to a deeply ingrained bilingual culture in either Sesotho to

English or English to Sesotho and the level of coverage differs and most of these dictionaries are targeted for second language speakers in the first instance and/or the general audience. Also, unlike trends in major languages of the world, Sesotho lexicography has been mainly a one-man driven project even though a lot of consultation and research about the content and semantics of words is conducted.

The latest two dictionaries Motsapi (2015) and Sesotho National Lexicography (2018) for the foundation phase of CAPS (the adopted methodology in South African schools) are the first two attempts at a Sesotho dictionary group work and it can only be hoped that team efforts will become the order of the day in the 21st Century. Disappointingly, though the teams are made up of scholars working at it for lexicographic sake and the development of Sesotho, the dictionaries are still bilingual and mono-directional from English to Sesotho! The big question is when will mother tongue speakers of the language be given first preference in Sesotho in lexicographic efforts of national proportions? Notwithstanding the bilingual nature of these latest dictionaries, Motsapi (2015) is a good example of the result of what happens when speakers of a language and language specialists develop a dictionary. The dictionary may have a few typographical errors and a few mistranslations, however, on a larger scale, the translation is of good quality. Nonetheless, the requirement for a monolingual South African dictionary, for a school learner's dictionary and for a monolingual online dictionary remains.

In terms of categorisation, these early products were general translation dictionaries. The compilation processes of these dictionaries are not comprehensively recorded or theorised. The compilers relied on the spoken word for the data collection and as a result, in some instances this resulted in mistranslations or misspelling of some words. De Kind et al. (2012:177) have observed that, as a result of indigenous terms being used to denote foreign concepts, "Their meaning is not just extended [in translation dictionaries]; their original

meaning [...] is denied and abandoned for the foreign concept. This becomes especially clear when comparing these terms to other Bantu languages or to the Proto-Bantu reconstruction." Aside from the challenges of translation, the Sesotho dictionaries produced then are building blocks for Sesotho lexicography. Even though there were no guidelines for compilation, that is metalexicography, these dictionaries can be studied and provide some good direction in terms of the dos and don'ts of writing dictionaries.

It was only in 1994 that the first Sesotho general purpose, monolingual dictionary was produced by a Sesotho mother language speaker, titled *Sehlalosi: Sesotho Cultural Dictionary* (Matšela, 1994). This dictionary was also produced in Lesotho and written in Lesotho spelling rules. However, it was more encyclopaedic. Two more monolingual dictionaries were produced, one in 1997 and another in 2005. Pitso (1997) published the former, *Khetsi ea Sesotho,* also in Lesotho orthography. Strictly speaking, *Khetsi ea Sesotho* is a thesaurus. The latter, *Sethantšo sa Sesotho* followed in 2005 by Hlalele, a priest who took 40 years to complete and publish it. It also was presented in Lesotho spelling convention. Though still usable by other users, the +9561 lemmata-strong, general-purpose compilation, targets users in the secondary and high school phase, has adopted the infamous, phonemic alphabet arrangement and is in paper format. Historically, paper dictionaries have formed the basis of electronic and online dictionaries. To date research does not show any activity in expanding it to include modern vocabulary resulting from advancing technologies or making an online resource.

Consequently, there are only two general monolingual dictionaries, which were produced in Lesotho (Matŝela 1991, Hlalele 2005). In the past, researchers have stated that the dearth in monolingual indigenous or previously marginalised languages dictionaries is due to shortage of sponsorship to undertake such lexicography projects (Nkomo and Wababa, 2013). The interest to develop a Sesotho monolingual learner's dictionary was also, in the process, to

identify micro and macro factors that affect the production of South African languages monolingual dictionaries, which may have deterred the progress and growth of Sesotho lexicography in the South African case. A survey of Sesotho dictionaries reveals that not only are there no monolingual learner's dictionaries but also there are no monolingual learners' online dictionaries.

Moeketsi (1991) also developed a multilingual dictionary in two different directions- English-Southern Sotho-Afrikaans and Afrikaans-Suid Sotho-English. The directions hereof speak of the target audience of this dictionary, Second Language (L2) Afrikaans and English speakers who are learners of Sesotho though this does not mean the Basotho cannot use it. This dictionary is basically a rudimentary wordlist with occasional Parts of Speech (POS) provided and is a quick ready resource for all users. In size it is consists of 122 and 123 pages in each direction. Each entry is in all three languages and starts with either English or Afrikaans depending on the direction.

The lexicographic needs of Sesotho have expanded due to its status as an official national language in South Africa and Lesotho. The use of Sesotho in the sectors of government, education and media renders it essential for the language to have reference resources to ensure the accuracy of the concepts it represents. Without a source of reference, language users are left to rely on intuition and speculations, which can lead to inconsistencies within a language in the absence of a standardised resource.

During the early initiation of Sesotho bilingual and trilingual dictionaries, translation dictionaries were resources that settlers compiled to bridge communication lines between themselves and the native South Africans (Kiango 2000, Ngomo and Wababa 2013). Bilingual and trilingual Sesotho dictionaries are important for language relations. However, this research holds that a monolingual dictionary with semantic, lexical, and pragmatic information can be

of great benefit to the first language speakers and learners of the language, and additionally provide a reference for translation dictionaries.

In summary, as regards paper dictionaries, the two surveys of Sesotho lexicographic history – the author and Setaka and Prinsloo (2020) - also show the following analysis: 16 paper dictionaries were produced and they can be further divided according to type as 1 thesaurus-cum-LSPD (language for special purpose dictionary), 14 bilingual dictionaries involving both Sesotho and English in either direction and 2 monolingual general dictionaries and 1 multilingual dictionary in Sesotho, Afrikaans and English. Bilingual dictionaries in English-Sesotho direction accounts for 7 dictionaries and Sesotho-English 3 only. All in all Sesotho lexicography is varied though there is a dearth of monolingual dictionaries, online dictionaries and as observed by Kiango (2000:10) a paucity of L1 lexicographers and institutions offering lexicography as a discipline and so producing experts in the field. This situation also obtains in both Lesotho and South Africa where Sesotho is spoken and is hampering speedier or upgraded work on Sesotho lexicography.

There has been migration into online dictionary production since 2006 with the publication of *Bukantswe Online v.2*, which broadened the typography of Sesotho lexicography. Online dictionaries, which are different editions of the same dictionary by the same author total 6 and 2 are in Sesotho-English direction and 4 in English-Sesotho. However, looking at the types and structure of dictionaries, Sesotho lexicography content can advance in the 21st century and moving forward the need for language resources in education, for example and for the language to take advantage of the human language technological developments of the current times can be attained. Gouws (1996: 99) adds that, “Although the South African languages have a comprehensive collection of dictionaries, it is not hard to detect imbalances between languages, types of dictionaries and lexicographic systems”.

Kiango (2000: 5) adds on to the above observation that, "[…] the tendency of the current Bantu lexicographers to perpetuate the tradition of compiling monolingual and bilingual dictionaries for foreign learners as their main dictionary projects, will not push the Bantu lexicographer one step further from where the pioneer lexicographers have left." Although Kiango's observations were made from a West African lexicography perspective, Sesotho lexicography is subject to the same deficiencies. When compiling a dictionary, the compiler should identify a specific problem under each component of the language. Kiango (2000: 10) lists factors contributing to problems facing African Languages lexicography and these include: a) that most languages are used in oral communication only, b) more than 90% of African Languages lack linguistic descriptions, c) there is a tendency for these languages to have many variants of the same lexical items, and d) there is a paucity of lexicographers and institutions dealing with lexicography. Considering the linguistic situation mentioned above, there is a need to involve linguists in developing Sesotho dictionaries to eliminate structural and meaning errors such as those found in early dictionaries. To reiterate, the latest Sesotho dictionary, which is a bilingual, word-based print dictionary the *English – Sesotho Bilingual Dictionary* (Motsapi, 2015) is an example of the result of what happens when speakers of a language and language specialists develop a dictionary. Therefore, it is time for Sesotho lexicographers to develop monolingual dictionaries; dictionaries for school learners and online monolingual dictionaries to stay on par with latest language and technology developments.

1.1.3. Corpus Linguistics

Previously, methods that were used to compile dictionaries used streams of collecting data from individual speakers through verbal conversation (Mabille and Dieterlen, 1979). However, with the advent of Corpus linguistics, the most dependable method of data collection is a compiled written and spoken corpus from varied text/sources. Due to time restraints and costs, this research used written texts in the main (c.f. Chapter 3).

Using Corpus Linguistics methodology towards the development of the dictionary was also identified as a low-cost strategy to accomplish the creation of an electronic dictionary. The methodology was also observed from how corpora were used in the design of the Northern Sotho and Setswana dictionaries namely *Pukuntšu ya Sesotho sa Leboa – Seisimane* (2003) and *Tlhalosi ya Medi ya Setswana* (2012). For instance, in the case of *Tlhalosi ya Medi ya Setswana*, Otlogetswe (2013: 535) states that "Frequency analyses were conducted to generate a dictionary headword list while word concordances were extracted to study how different words behave in context to aid sense discrimination." The work in the 2012 Setswana monolingual dictionary demonstrates how analyses in corpus studies show structural qualities of a language that are required for a linguistically representative monolingual dictionary. It was distinguished through the research that corpora are effective resources of language because a corpus provides lexical information and the structure of a language can be studied from it.

On the other hand, Northern Sotho has an increasingly growing corpus, although described as imbalanced and not representative of all the dialects of the region (Mojela, 2013). Notwithstanding, Northern Sotho used its available corpus to develop *Pukuntšu ya Sesotho sa Leboa – Seisimane* (2003), and for typographical expansion, it employed technological methods to present and make the dictionary available on the *World Wide Web*. This came with the introduction of *TshwaneLex* dictionary software. And this is the software used for the creation of the prototype monolingual Sesotho Learner's dictionary (Appendix 1).

1.2 Statement of the Problem

At the initial start of this study in 2014, one could not find the following types of dictionaries in Sesotho available to purchase in South Africa: learner's dictionaries, monolingual print or online dictionaries, or monolingual learner's dictionaries. It is for this reason that this research has initiated the investigation reported herein to close this gap. According to De Kind et al.

(2012:116), lexicography was introduced to African Languages about 360 years ago with the *Vocabularium, Latinum, Hispanicum, e Congense* in 1652 by Van Gheel (De Kind et al., 2012) as the precursor. Sesotho lexicography is over 160 years old. That Sesotho has a well-standing tradition of practical bilingual, print or paper lexicography even though the target end-user was, mainly, the Second Language (L2) audience is a moot point. There remains to date a dearth of suitable dictionaries for L1 speakers of Sesotho. Sesotho dictionaries started being developed in 1852, as discussed in Section 1.1.2. above. Sesotho was the first South African South Eastern Bantu language to be reduced to writing with all other sister languages following in its footsteps. Other Sotho family languages such as Setswana and Sesotho sa Leboa have moved from the bilingual, translational, paper dictionaries and ventured into making monolingual online dictionaries. However, Sesotho, despite its almost +160 years of lexicographic tradition, has not made serious developments. As indicated above, within the period of Sesotho dictionary-making, there have only been three monolingual lexicography products: Matšela's *Sehlalosi: Sesotho Cultural Dictionary* of 1994, Pitso's thesaurus, *Khetsi ea Sesotho* of 1997 and Hlalele's *Sethantšo sa Sesotho* of 2005.

Matšela's (1994) dictionary states its type and target in the title i.e., it is a cultural dictionary, and this type of dictionary can be used by all speakers of the language for cultural reference. It is an SLP, that is a special language dictionary instead of a general one. Pitso's (1997) on the other hand, is a thesaurus and thus restricted in the type of information one can find in it.

Hlalele's (2005) dictionary seems to cover a larger spectrum because reviewers state that, "It is also useful to students, teachers, lecturers and university students who are doing African languages and literature and libraries all over the Southern African Development Community (SADC) countries." Motjope-Mokhali et al. (2020:262). Further, the trio (ibid. p.

262) aver that it offers detailed information regarding word categories, such as nouns and verbs, morpheme boundaries are separated by using decimal points, word usage is contextualised and other lexicographic labels such as subject area are also provided. It is not as limited as the other two monolingual dictionaries, Matšela (1994) and Pitso (1997). However, Hlalele's (2005) dictionary is written in Lesotho orthography, therefore, it causes confusion for modern-day South African school learners who also have to grapple with different orthographic conventions. Lastly, finding this dictionary has been like the proverbial needle in a haystack as it has not been available for an exceptionally long time in the market, possibly, due to publication glitches.

Secondly, it is noticeable that there have been developments as regards Sesotho entering the digital dictionary scene with the introduction of the two volumes of *Bukantswe Online*. However, the online volumes are also bilingual, and their microstructures do not scratch the surface of modern lexicography. They are rather rudimentary in the modern era in relation to what a microstructure of an entry explained earlier according by Hartmann (2012) and Imbayarwo (2008) should entail, and in comparison, with Mabille-Dieterlen (1852-1974) of the ancient era and Hlalele (2005). Though the bilingual dictionaries projects have played their role in Sesotho for the purpose of their target users who are initially, foreigners learning Sesotho and they can be said to be useful for translation purposes in general, they have also been useful to an appreciable degree for sufficiently literate mother-tongue speakers. However, for the learners who are in a South African school and taking Sesotho as a school subject, it is imperative that a monolingual dictionary be available.

Thirdly, the methods and tools that were used to develop the existing Sesotho dictionaries were not documented as was the case in many ancient efforts of other languages as theoretical lexicography had not been born yet. The little literature available predicates that

the developers of early Sesotho dictionaries, who were missionaries and colonial administrators (Kiango, 2000), gathered lexicographic information from the speakers. However, the literature does not detail the procedures of how the data that was collected from the speakers was documented and verified. Concerted to the little articulation of the data collection methods of early and existing dictionaries is also the scarcity of Sesotho theoretical lexicography. With about 160 years of dictionary making in the language, there are very few theoretical accounts on Sesotho lexicography that could be traced from lexicography and linguistics journals. Atkins, Clear and Ostler (1992: 4) briefly states that theoretical lexicography is "a body of theory related to lexicography" and Sesotho is also trailing behind on this aspect. Additionally, the methodologies that informed the compilation of most of the existing Sesotho dictionaries are inarticulate in the discipline of lexicography.

Kiango (2000) identified that there are problems with African Languages lexicography, which are related to factors of interlanguage, the internal structure of the language, that is, phonological, morphological, syntactic and semantic structure. Another problem has to do with extralinguistic factors. These are non-linguist factors which are considered important components in dictionary making, that is, goals of compiling a dictionary, the type of dictionary, a native speakers' dictionary, a learner's dictionary, and advanced learner's dictionary, or a monolingual dictionary (Kiango, 2000: 1).

To date, there are only two general monolingual dictionaries, which were produced in Lesotho (Matŝela 1991, Hlalele 2005). In the past, researchers have stated that the dearth in monolingual indigenous or previously marginalised languages dictionaries is due to shortage of sponsorship to undertake such lexicography projects (Nkomo and Wababa, 2013).

1.3 Aims, goals, rationale and research questions

1.3.1. Aim of research

The aim of this study was to develop a monolingual Sesotho Learner's dictionary using Corpus Linguistics approaches, specifically, the corpus-based approach.

1.3.2. Goal(s) of research

The goal is to develop a monolingual Sesotho learner's dictionary of +5000 lemmas, a realistic size for learners, to be used primarily by school learners as a reference resource, and to ultimately make it available online. However, a dictionary of that size is too big a project for a Master's research and so in order to achieve this goal, the intended work was divided into three phases – development the corpus and the prototype, populating the macrostructure of the rest of the +4700 remaining lemmas of the dictionary and finally customizing it to an online resource. As a result, the immediate goal is to process the first phase, developing a corpus and prototype that will serve as a model of the entire effort hence only A-B stretches are covered.

1.3.3. Rationale

This research was driven by the understanding that a dictionary is a basic resource to learning and a monolingual one is critical in the learning of Sesotho as a subject. Language learners need a dictionary to support them in developing language and learning skills. Secondly, the world at large has moved to advanced human language technologies and so while the paper dictionary is useful and will remain so for other users, an online dictionary is unavoidable and in keeping with the recent and future times. Further, this research studied other Sesotho dictionaries and the corpus texts themselves and identified serious spelling inconsistencies across and so the availability of such a resource will promote the standardized writing

conventions of South African Sesotho while not precluding Lesotho orthography as the differences in writing between the two is minimal and generally understood in both countries. The dictionary sought to use the current standard Sesotho spelling as advocated by PanSALB to address the inconsistencies discovered.

1.3.4. Research Questions

By setting on to develop this dictionary, the umbrella question, which helped pave the way for the methods that would be used to develop the dictionary and the analysis thereof was:

(i) What are the factors of consideration when developing a dictionary with these specifications? This question was further broken down into the following:

(ii) What kind of data and tools can be used to develop a monolingual online learner's dictionary?

(iii) What kind of information can be marked up for a monolingual dictionary that can be useful for language learning?

1.4 Framework and Research Design

As Kiango (2000) identified, there are problems with African Languages lexicography, which are related to factors of interlanguage, the internal structures of the language, that is, phonological, morphological, syntactic and semantic structures. Another problem has to do with extra-language factors. These are non-linguist factors which are considered important components in dictionary making, that is, goals of compiling a dictionary, the type of dictionary, a native speakers' dictionary, a learners' dictionary, and advanced learners' dictionary, or a monolingual dictionary (Kiango, 2000: 1). Wiegand (1983) adds that,

> Lexicographical activities result in reference work which can be classified according to different types. All types of works made with the aim of providing not only, but above all, information on linguistics expressions should be classified as linguistic lexicography. They would include at least the following types: dictionaries of language, glossaries, concordances and word indexes. [...] We can characterize the subject area linguistic lexicography, as follows: Linguistic lexicography is scientific practice aimed at producing reference works on language, in particular, dictionaries of language. Lexicographical activity has recourse to the results, methods and theories of various academic disciplines according to the type of reference work being produced. (Wiegand, 1983: 16)

Notwithstanding, lexicography can be applied from a semasiological approach or onomasiological approach. The Semasiological approach to lexicography "[…] is an approach in which the linguistic form is the object of study" (Warbuton, 2015: 374). From the perspective of information or reference science, lexicography conveys data using a semasiological approach. Rey (1995) maintains that lexicography concerns itself with the function and the behaviour of words in society. It proceeds semasiologically, i.e. it takes signs and studies their meaning and the concepts they stand for. The onomasiological approach on the other hand, is an approach that is more related to the terminology subject area. This approach deals with the description of concepts by assigning meaning, according to General Theory. The research study at hand deals with lexicography from both perspectives.

There are other principles and methods regarding arrangement of lemmata in a dictionary and the one widely used method is the alphabetic arrangement, in particular, the Latin alphabet which is based on the Greek and the Greek is based on the first known alphabet, the Phoenician alphabet. The letters of the Greek alphabet, as a general rule, represent simple sounds. However, Sesotho abounds with complex sounds represented by digraphs and trigraphs and so posed serious challenges for earlier developers. The result is that the words with complex-initial sounds inevitably represented by several letters of the alphabet such as /th; tsh; tjh/ etc. were classified separately from the simple-initial consonant words. As such a word such as *tumo* 'fame' would come before a word such as *thaba* 'mountain'. The alphabet stretches dealt

with simple-initial consonant words followed by a vowel first and then provide a different article stretch of complex-initial consonants followed by vowels. The problem with this arrangement is that users not familiar with the phonemic alphabet of the language as opposed to the phonetic one, have great difficulty locating words in these dictionaries. To the researcher's knowledge these differences are only taught at university level to linguistics students. This was the nature of early dictionaries including Hlalele's (2005). Notwithstanding, alphabetic sorting of lemmata relates back to the semasiological approach and its greatest advantage is that it enables users to look up words in a dictionary in an ordered manner, and this is the arrangement that the early dictionaries have all used.

Another factor regarding arrangement of lemmata is whether it uses full words or stems. Earlier dictionaries preferred the stem-approach especially for nouns, verbs and qualificatives and the derivatives arising therefrom are listed as sub-lemmas. It has been argued that this method is problematic when it comes to indigenous South African Languages because these languages are agglutinative in structure. Kiango (1992: 14-15) states that, "as a remedy to this problem, lexicographers 'adopted' the practice of listing nouns with prefixes, and listing verbs and adjectives by stems. This system also solved the problem of grouping related items together." As regards the arrangement of words in a dictionary, the bottom line according to Imbayarwo (2008:3) is that dictionary 'data should be structured in such manner as to facilitate optimal and successful retrieval of desired information from the lexicon.

Concords, also known as agreement markers - which establish agreement between the subject or object noun with the pronoun, the qualificative and verbs - and tense markers have the highest frequency than the lexical and fixed categories thus confirming their centrality and import in Sesotho morphology and that of its sister languages - Setswana, North Sotho and Nguni - and the Bantu family at large. Though they are not parts of speech by themselves, they

are an important part of any part of speech they prefix to and so need to be included in any dictionary. The designed dictionary has opted for full word lemmas in the challenging three categories – nouns, verbs and qualificatives.

The lexicographic method and all other methods of processing the data for this dictionary are both quantitative and qualitative, which are both accommodated in the corpus-based approach. The quantitative part involved sampling texts for the corpus, generating a wordlist and studying word frequency counts, and selecting lemmas for the dictionary – it was applied on the macrostructure of the dictionary. The qualitative methods on the other hand, applied to the microstructure of the dictionary where the structure of lemmas was studied and aspects such as appropriability and spelling were checked; the contexts of the lemmas was studied from the corpus to see the number of senses or POSs they occupied in order to assign them the relevant definitions, and for extraction of usage examples. Further, the compiler is a 1st Language Speaker of Sesotho, and the knowledge of the language was also applied in assigning examples and meaning and assessing the spelling of the words etc. For the dictionary microstructure, the semasiological approach to lexicography was applied.

Finally, the *TLEX* suite used for corpus compilation, developing wordlists and their frequency counts, using frequency counts to select lemmata and the corpus for meaning and example sentences also embraces both qualitative and quantitative approaches.

1.5 Results: their significance, practicability, and generalisability

The function of any dictionary is to serve as a reference resource. Dictionaries reveal the status of words, meanings, and usage. Hanks (2010: 581) observes that "words have the potential to make meanings when put into context – and they are associated with particular sets of syntagmatic patterns, which can be discussed through painstaking corpus analysis." Hanks

(2010) identifies that there are two major kinds of monolingual dictionaries; these are the tradition-based and the synchronic dictionaries. The tradition-based dictionary reports of history and etymology, i.e., it indicates how the word has evolved over a period to what it has come to mean in the present day – it is diachronic. On the other hand, "the focus in a synchronic dictionary is on reporting conventional meanings and use, rather than on historical and etymological research" (Hanks, 2010: 582).

A prototype dictionary dubbed *Sehlalosi sa Baithuti sa Sesotho* has been developed as the result of this research and presented here, it has both synchronic and diachronic attributes. Etymological information is provided for lemmas were necessary and the tool that was used makes provision for the addition of etymology of each lemma in the article structure. It is also synchronic in that it provides lemma definitions and usage examples from various types of texts from varying periods of writing. The research report presents the prototype monolingual Sesotho Learner's dictionary (letters A – B), although the initial and ultimate goal was to undertake the process of dictionary making and develop a fully flashed online dictionary of 5000 words.

When the research was initiated, there was no accessible Sesotho corpus readily available and therefore, a corpus of 101 359 words was created, which comprises of two Sesotho orthographies. The corpus is annotated at header level, and the variety of text samples it covers has been organized according to genre and subject specific glossaries could be developed from it.

1.6 Scope

The study looked at Sesotho lexicography, with a focus of contributing to Sesotho monolingual lexicography. Additionally, the study explores Corpus Linguistics because of having used a

corpus as a tool in the dictionary development. Corpus Linguistics has two approaches to it, corpus-based and corpus-driven. This research adapted the corpus-based approach. Although the corpus used for the dictionary was collected from South African and Lesotho Sesotho texts, the dictionary was developed with the South African school audience in mind. Thus the inclusion of school learners' texts in the sampling. The school texts sampled are for Grade 10, however, the dictionary could be used by all High school grades. More school texts used in schools from other provinces where Sesotho is taught, will be added on the corpus for expansion for future projects. The prototype presented and analysed herein, the immediate goal of this research, includes +250 lemmata.

1.7 Chapter Outline

Chapter 1:

This dissertation consists of five Chapters, and the aim of Chapter 1 was to give a background of the factors that informed the initiation of this research and its significance. To lay out the focus of the study and point out what the study achieved, and the methodologies used to achieve the aim of working towards the development of a monolingual online Sesotho Learner's dictionary.

Chapter 2:

Chapter 2 draws from the already existing research in the fields of general lexicography and Corpus Linguistics (CL). It has reviewed the literature with respect to the historical background of the language, its morphology and draws from the already existing research in the fields of general Lexicography and Corpus Linguistics (CL). It provides a brief history of Sesotho and its development over the years looking at its linguistic aspects and the structure of parts of speech (POS), its two orthographies living side-by-side since the early 1960s and variations

thereof, the Sesotho scope of lexicography and its traditions and dictionary-making theories in general. The latter has indicated that on the main, bilingual dictionaries have been the preferred choice and that existing online dictionaries - which are still bilingual and monodirectional - are still rudimentary in Sesotho lexicography. Further, the culture of monolingual dictionaries in Sesotho is also still at its infancy as of over 20 dictionaries surveyed, only three are monolingual and only one is a general dictionary. It is the aim of this study to address these deficiencies to meet the current social needs of Sesotho mother language speakers. As lexicography deals with words, a review of the morphology of Sesotho was undertaken to inform the decisions on the compilation of the envisaged learner's dictionary.

The literature under review looks at the gaps that Sesotho has not claimed in the fields of Lexicography and Corpus Linguistics, and further studies the different conventions and guidelines that have informed this research towards the advancement of Sesotho in dictionary making, metalexicography and Corpus Linguistics. The semasiology and onomasiological approaches of lexicography will be discussed. One of the factors in the chapter is the contribution of Corpus Linguistics in lexicography, which emanate from the literature on the projects that have used either the corpus-based or corpus-driven approach before. The literature supports the decision made to utilise the corpus-based approach in this study. The next chapter will outline the methodologies and methods that will be applied in Chapter 4.

Chapter 3:

Chapter 3 is mainly on the research design aspect, that is, on the methodologies used in this research. This chapter is divided into two main categories: corpus compilation and the actual design of the dictionary. The main tool, the corpus, is broadly described under the factors that inform the nature of a standard corpus which are: size; machine-readability; the aspects of annotation and balance and representativeness and how it was processed and used as data for

the envisaged dictionary. The processes of collection, sampling and sorting of the texts are also illuminated with both the corpus-driven and the corpus-based approaches juggled according to the needs of the research. Secondly, it discusses the methods used in the actual compilation of this dictionary and justifies the semasiological over the onomasiological approach as it was deemed fit for this research. Additionally, the chapter makes reference to the various dictionary making tools and draws focus to the properties of the chosen tool, *TLex,* which is a multifunctional dictionary making suite with additional modules for corpus compilation and a concordance all in one package amongst many others and so making its use most ideal for the development of the dictionary. Last and most importantly, both methodologies for corpus compilation and the development of the dictionary are both quantitative and qualitative in nature. All these three are inseparably connected to yield the 250 lemma-strong prototype product described in the next chapter.

Chapter 4:

Chapter 4 analyses the data collected using the methodologies outlined in the previous chapter for both data compilation and processing data into dictionary format ending up with the resultant prototype dictionary as per the aims of the research. The Chapter demonstrates amongst others, how much of each subject field corpus was covered to render the corpus as inclusive and general as possible, how the wordlist was generated using the concordance tool *tlCorpus*, which range of words were selected and how the wordlist was exported to *TLex.* It provides illustrations of the processes undertaken towards the creation of the dictionary, especially developing both the macro and microstructure of the dictionary, and clearly delineating their features such as POS distribution, marking, pronunciation varieties and indicating homophones and homonyms where applicable over and above syntactico-semantic information that forms the articles of each listed lemma. The steps of adding lemmas,

converting the *TLex* default language to Sesotho so that the lemma properties such as POS and cross-reference types could be in Sesotho are also demonstrated. The challenges and decisions that were made along the way based on the discoveries of how the dictionary tool works are also treated in the chapter.

Chapter 5:

Chapter 5 discusses each of the main results obtained in relation to the questions raised and aims outlined in chapter 1. The discussion compares the results of this study to other research efforts and/or projects that used similar methods as regards their significance, practicability, and generalisability. It also puts forth the implications, limitations, and challenges of the study and makes recommendations for future work. In conclusion, notwithstanding the unavailability of most of the surveyed dictionaries from libraries in the OPAC system, and paucity in Sesotho corpora at the initiation of this research, the study has succeeded in developing a prototype for the envisaged ultimate goal of this research and will in the future include features that are still lacking to realise the desired product. The prototype dictionary is annexed at the end.

CHAPTER 2: LITERATURE REVIEW

Introduction

The purpose of this research is to design an online, corpus-based monolingual Sesotho learner's dictionary. The focus of this chapter is to provide the historical background of Sesotho first and then a broad review of the morphology of Sesotho words, general history and development of African Lexicography, in particular, that of Sesotho and the conventions of dictionary-making. The section on morphology provides a classification and description of Sesotho words as they are a *sine qua non* for dictionary development while lexicographic review not only contextualises the study further, but also informs on the issues entailing dictionary making.

2.1 Historical background of Sesotho

2.1.1. Overview of Sesotho

Sesotho which originates from the Niger-Kordofanian family is a language of southern Africa, which belongs to the Niger-Congo Bantu languages family, a branch of the Niger-Kordofanian. The generalised tree in 2.1. below from researchgate.net traces the origin of Sesotho from the Niger-Congo large phylum and shows its relationship with other languages in the Niger-Congo family. The Niger-Congo, also known as sub-Saharan languages, is an enormous branch whose languages are found throughout South and Central Africa and in most of West Africa below the Sahara. This family divides into three branches: the Mande, Atlantic and Kordofanian. Volta-Congo which stems from the Atlantic, branches into four sub-families Kru, Benue-Congo, Kwa and North Volta. Bantoid is the daughter of Benue-Congo whose grandchild is Bantu. The latter includes the largest language family of sub-Saharan Africa, the Bantu languages, comprising hundreds of languages. About 120 million people speak these

languages. The important branches in Sesotho lineage are coloured in green in the diagram below.

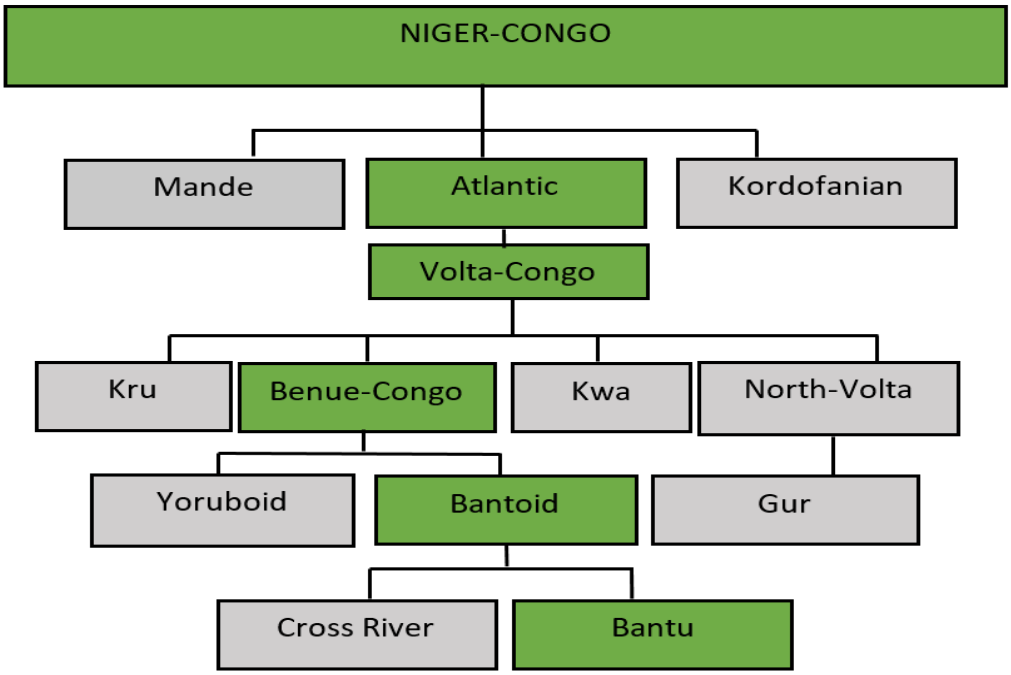


Figure 2.1: *Niger-Congo family tree*

Significant Bantu languages include the Indigenous languages of South Africa, namely, Xitsonga, Tshivenda (also known as Tsonga and Venda respectively), Nguni and Sotho. Nguni includes IsiZulu, IsiXhosa, IsiNdebele and Siswati whilst the latter comprises Setswana (previously known as West Sotho in the Republic of South Africa), which is spoken in Botswana and also mainly in the North West Province of South Africa and Thaba Nchu in the Free State; Sesotho sa Leboa (also known as North Sotho) which is spoken mainly in the Limpopo Province, previously Northern Province, and finally Sesotho (also known as either Southern or South Sotho in early literature).

The organogram in Fig.2.2. shows the place of Sesotho, a member of the Sotho group in the South Eastern Bantu family.

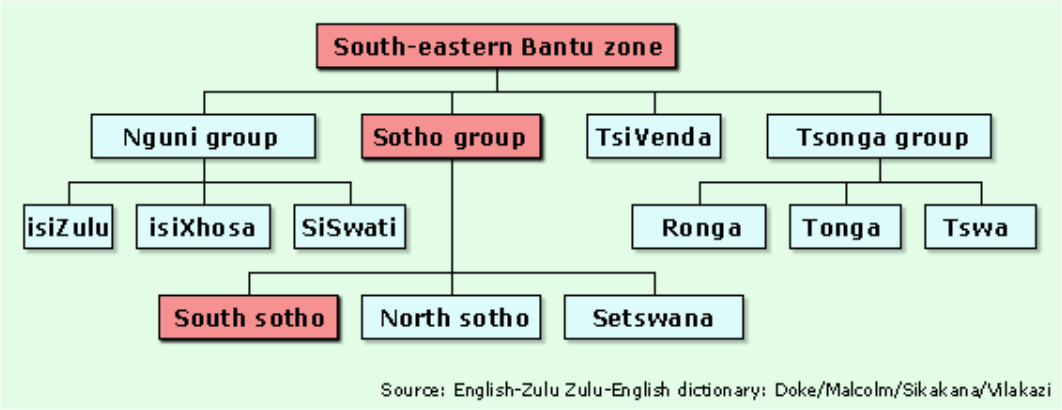


Figure 2.2: *South-eastern Bantu languages*

The five major dialects of Sesotho are, Sekgolokwe, Setlokwa, Sekwena, Serotse and Sephuthing and the written variety is mainly based on Sekwena, the standard official dialect in both Lesotho and South Africa. Sesotho speakers are found in Lesotho and parts of South Africa in the Free State, Gauteng, Eastern Cape and North West provinces. Statistics (Census 2011) indicate that ± 7.6% South Africans and 86% of the Lesotho population are First Language (L1) speakers of Sesotho. Sesotho falls in the same language family as Northern Sotho and Setswana by sub-categorisation. It shares phonological, morphological, syntactic, and semantic similarities between these two variations. Syntactically, Sesotho is a subject-verb-object (SVO) language meaning that this is the order of these three elements in a sentence in the non-emphatic or regular mode of speaking or writing. Morphologically, Sesotho and its immediate relatives makes use of affixes - one or more - which mark agreement between subjects and verbs, verbs and objects and mark number, that is, singularity and plurality. Sotho languages are agglutinative, with clear boundaries between the stem/root and affixes and use a hybrid orthography of the conjunctive and disjunctive writing system, the former for nouns and

the latter for verbs, qualificatives and some adverbs and they follow similar rules in word-formation.

Figure 2.3: *Areas where Sesotho is Spoken*

The 1800s were a time during which missionaries and colonial administrators were infiltrating southern Africa. As a result, there was significant production of bilingual dictionaries during the period to meet the social needs of the missionaries and colonial administrators (Kiango, 2000). The mission, which was to learn the native languages to inverse teach the natives English and Afrikaans, was accomplished therefore, the 21st century and going forward calls for dictionaries that will meet the current and future needs of society. English is a dominant language in South Africa; speakers of the previously marginalised language are bombarded with English from Grade R, at age 5 – 7 years. At this stage, children have not developed strong cognisance of this their mother languages yet the way that the public education system is, for years, learners have to learn a second language regardless. A bilingual dictionary would, of course, serve useful to these learners only if they had sufficient vocabulary in their mother language.

Sesotho is an official language in Lesotho, alongside English; it is used for national address and is a primary medium of instruction in education. In contrast, it is one of the 11 official languages of South Africa (The Constitution of South Africa, 1995), but as is the case with the rest of the other previously marginalised official languages, it is a secondary medium for national address and education. For example, government communication is presented in English and then translated into other languages by some government departments, municipalities, and entities. On the other hand, in school, technical subjects are taught in English and Afrikaans as the designated L1 or Home Languages (HL) and are the Medium of Instruction (MOI) in addition to being language subjects on their own. Sesotho and other indigenous languages are subjects but not MOI. This is the case even though 80% of the South African population is not L1 speakers of Afrikaans or English.

2.1.2. The Orthography of Sesotho

From the 1830s when the missionaries in Lesotho reduced Sesotho to writing, there was only one spelling convention in both Lesotho and the Union of South Africa. However, in 1961 when South Africa became a Republic, a new orthography for Sesotho was implemented in the republic. Since then, the two orthographies have co-existed. Figure 2.1. below captures these differences which involve the glides and a number of consonants. For example, in South Africa, diacritics are no longer used in writing, but Lesotho Sesotho continues to follow the first spelling that was adopted when the language was reduced to writing in the 1840s. The difference in the two variations is in the transcription, but the pronunciation of words remains the same. This information is noted because some could argue that if there are language learning resources published in Lesotho, there wouldn’t be a need for similar products to be developed for South Africa.

Sound Class	RSA	RSA Example	Lesotho	Lesotho example	Gloss
Glides	w/o	wena	ue	Uena	‘you’
	Y	yena	ee	Eena	‘that one’
Consonants	D	dieta	l	Lieta	‘shoes’
	Kg	kgotso	kh	Khotso	‘peace’
	Kh	khabetjhe	k’h	k’habeche	‘cabbage’
	Fj	le**fj**wa	fsh	le**fsh**wa	‘get paid’
	Mm	mme	‘m	‘me	‘mother’
	Nn	nnete	‘n	‘nete	‘truth’
	Nng	nngwe	‘ng	‘ngoe	‘one’
	Nny	nyantsha	‘ny	‘nyantša	‘breastfeed me’
	Pjh	m**pjh**e	psh	m**psh**e	‘ostrich’
	Tjh	tjhelete	ch	Chelete	‘money’
	Tsh	tshimo	tš	Tšimo	‘field’

Table 2.1: *RSA and Lesotho orthography differences*

To further illustrate, the phonology of Sesotho in both countries is distinct in the case of vowels and the use of the letters /d/ and /l/. In Lesotho they use all vowels to form a syllable with /l/ but when /l/ is preceded by either /i/ or /u/ it is articulated as /d/. In RSA, only /i/ and /u/ form a syllable with /d/ and the remaining vowels /a, e, e, o and o / always precede /l/ and never /d/ in word-formations. Sesotho uses a seven-vowel system. These are: [a], [e], [ɛ], [i], [o], [ɔ], [u]. These vowels are spelled the same as /a/, /e/, /e/, /i/, /o/, /o/, /u/. In the case of the **e’s** and **o’s**, other letters or words in the context of these letters can help to distinguish the sound that a vowel is representing, which also helps to determine the meaning of the word, particularly when looking at homographs such as, *l**o**na* “you all” vs. *l**o**na* “it”, and *emere* “bucket” vs. *emere* “pregnant”. The mid-high /o/ is spelt as /u/ in Lesotho orthography because they are similar in sound; as a result, there are instances in Lesotho orthography where one finds *u se ke wa tsuba* instead of the South African *o se ke wa tsuba* “do not smoke” where /u/ is /o/ for [ɔ].

Of the 44 letters of the Sesotho alphabet, 35 are consonants. Sesotho has consonants represented by single, double and treble alphabetic letters, for example, /l/, /tl/ and /tlh/ or /s/, /sh/ and /tsh/. However, Lesotho orthography still uses diacritics in what should be triple

consonant sounds, for example, instead of ***tsh****aba* it uses ***tš****aba* “be afraid”. To reiterate, there are instances in the spelling where Lesotho Sesotho uses the letter /l/ for the sound [d] when it comes before the vowels /i/ or /u/. For example, *lihela* “drop” or *lulela* “sit for”, but when /l/ comes before /e/, /a/ or /o/ the pronunciation remains /l/.

2.1.3. The semantic theory of Sesotho

The Semantic theory of a language is a part of its linguistic description. It assigns semantic contents to expressions of a language. With the new formation of Sesotho representative terms or words in place of synthetically modified loanwords, the definitions for these words will need to be in a manner that the speakers, and primarily, the school learners taking Sesotho as a school subject, will be able to understand and be able to use the words in producing new sentences. Katz and Fodor (1963:171) state that, “[...] what qualifies one as a fluent speaker is not the ability to imitate previously heard sentences but rather the ability to produce and understand sentences never before encountered.” This is what language users, in general, using a verbal or written form of language have to undergo in everyday communication - they hear new sentences from media, school textbooks, community members and produce new ones themselves. Therefore, it is necessary that they understand these new sentences, but without a conceptualisation and annotation in mind, i.e. knowledge of the meanings of lexical items that make up utterances, there cannot be a complete understanding and succinct application of these lexical items in forming new sentences for interpreting the existing ones.

A monolingual dictionary containing Sesotho rule-related terminology will address factors such as ambiguity. Although there is not much scholarship on Sesotho Semantic theories, on discussing semantic theories for English structure, Katz and Fodor (1963: 175) have a view that can be used to hypothesize Sesotho semantics about lexicography and terminology. They assert that:

> [...] a speaker can disambiguate parts of a sentence in terms of other parts and thereby determine the number of readings of a sentence. Thus, another facet of a speaker's semantic ability is that of determining the number of readings that a sentence has by exploiting semantic relations in the sentence to eliminate potential ambiguities.

Although their discussion is focused on sentences, this view is supportive of how lexicographic practice can draw out ambiguities at word level, for further disambiguation at sentence level. This can also be applied in translation and contribute to second language acquisition as well as language comprehension and performance. On the other hand, a sentence composed of words from two linguistic backgrounds, and with no source of reference at the learners' disposal, the learner cannot be expected to assume factors such as these. Additionally, this is connected to the importance of having Sesotho structural-based terms and words in the place of synthetically modified loanwords, as Sesotho is an affixing language with morphemes that carry meaning on their own before they are attached to roots or stems. Some of the loanwords found in Sesotho cannot be decomposed for meaning analysis or morphological analysis. It, therefore, becomes critical to understand the semantic theory of a language in lexicography. Moreover, with there not being much on Sesotho Semantic Theories, the corpus compiled herein is a useful tool for studying the semantic structure of Sesotho, which additionally adds credibility for meaning extraction for the single lexical items.

A dictionary provides information on parts and aspects of the lexicon of a language if it is monolingual. There are different parameters set out to classify the different types of dictionaries. However, these all emphasise the aspect of purpose and function of a dictionary. Dictionaries, regardless of the number of languages they may comprise are primarily classified as general (semasiological) or specialised (onomasiological). The typologies of general and specialised are what distinguish lexicography from terminology dictionaries, respectively, without overlooking that general dictionaries can contain aspects of lexicography and terminology. Further, these two typologies can be classified under the aspects of coverage,

finance, access structure, and target user (Podhajecka, 2004:2). The classifications are done following the features of a dictionary. Lehmann (2013) states that:

> The most common kind of dictionary is one that is devoted to the vocabulary of one language. Its macrostructure is an alphabetical list of [lemmas]. Its microstructure comprises various kinds of linguistic information on each [lemma]. Apart from the word-list, the typical dictionary contains a couple of other components, e.g. instructions for use and a list of abbreviations.

Assam (2017: 379) alludes to the case of the development of Gabonese mother tongue education dictionaries that, "In the field of lexicography, it has been stated that the development of lexicography and dictionary production would certainly contribute to mother tongue education implementation in Gabon." The lack of monolingual dictionaries for education for Gabonese language is equally true for Sesotho, and Assam (2017) proposes a plan for education dictionary production for mother language speakers – an organisational plan and a dictionary conceptualisation plan, which are mentioned as the two planning phases of a lexicographic process. "Dictionaries must first be conceived and then written. Planning is the design phase for creating both the text specifications and the way the whole text will be written" (Assam, 2017).

Assam's study makes appealing points for this research including that "In order to be a functional instrument, a planned dictionary has to fulfil the purpose of empowering the learners in their language acquisition endeavour" (Assam, 2017: 383). This is important to note so that the use of technical linguistic symbols like those used in phonetics could be eliminated from a learner's dictionary because as things stand in the Sesotho school grammar books, learners are not taught phonetic symbols at basic education phase which is particularly important information for a language such as Sesotho which has a large number of homophones. "The genuine purpose of such dictionaries is to transfer, by means of lexical data, information regarding the set of lexical items included as treatment units in order to ensure the linguistic

empowerment of the target user" (Assam, 2017: 383). The next section will review the use of Corpus Linguistics in lexicography because the lexical items that are used for the dictionary were harvested from a Sesotho corpus.

2.2 The morphology of Sesotho

The parts-of-speech Sesotho uses are divided into major and minor classes. The major classes are nouns, pronouns, verbs, and qualificatives. These lexical classes are the same in both regions where Sesotho is spoken. Therefore, since the corpus is made up of both Lesotho and South African written texts, there might be cases where a word such as 'noise' has two representations such as *molumo* and *modumo.* A bird's eye view of Sesotho morphological categories is shown in Table 2.2.

Open	**Hybrid**	**Closed**
Nouns	Pronouns	Conjunctives
Verbs	Ideophones	Clitics
Adverbs		Interjections
Qualificatives		

Table 2.2: *Sesotho Parts of Speech categories*

2.2.1. The Morphology of the noun

Doke and Mofokeng (1957) identified 18 noun classes which are distinguished by their noun prefixes as in table 2.3 below where all these prefixes are bolded. The table is arranged according to Meinhof (1932) where classes are paired as it is widely used in many grammar books according to Mchombo (1998:517). The grouping of nouns is also based on some general characteristics as per table although classes such as classes 3-4 are heterogenous. The rule of thumb seems to be the class prefix, particularly, plural ones when not sure of the class of a noun. Classes 1-10 go in pairs with the odd numbers being singular word forms and the even

numbers being the corresponding plurals of the odd classes. There are no classes 11-13 in Sesotho.

A comparison of Sesotho with its sister languages seems to suggest that these missing classes have not really all been lost, but some are absorbed into other classes thus leading to extra plurals as in class 6 (Demuth, 2000). Corbett (1991: 43) confirms this observation by explaining that Proto-Bantu had twenty-five classes although currently no Bantu language shows all the twenty-five as some have been lost or merged with other classes. Class 14 takes a plural with loan words, and it is a class of abstract words. Class 15 is for infinitives used as nouns when classes 16-18 are adverbials, specifically, locatives. These class prefixes attach to primitive/primary roots or stems of other lexical categories. Both the primary root/stem and the class prefixes are bound morphemes in nature as neither of them can form a complete word on their own

Class	**Characteristic**	**Example**		
		Singular	**Plural**	**Gloss**
1-2	Names of persons	**Mo**tho	**ba**tho	'person/people'
1a-2a	Proper Names Kinship Terms	**Ma**lome	**b**omalome	'uncle(s)'
3-4	Impersonal Nouns	**Mo**tse	**me**tse	'village(s)'
5-6	Natural Phenomena Body Parts Others	**le**tsatsi **le**tsoho **le**role	**ma**tsatsi **ma**tsoho **ma**role/**di**thole	'day(s)' 'hands' 'feather(s)'
7-8	Person with certain habits/characteristics Objects	**se**tsebi **seta**	ditsebi **di**eta	'expert(s)' 'shoe(s)'
9-10	Animals Deverbatives Loanwords/Adoptives	**n**tja **(n)**pono **Se**kotlolo	**di**ntja **di**pono **di**kotlolo	'dog(s)' 'vision(s)' 'dish(es)'
14-6	Abstract Collective	**bo**rena **bo**hoho	n/a **ma**hoho	'chieftainship' 'crust'
15	Infinitives used as nouns	**ho** utlwa	n/a	'to hear'
16-18	Locative Nouns	**ho**dimo **bo**siu	n/a	'above' 'at night/night time'

Table 2.3: *Noun classes*

Noun class prefixes also play a critical part in the morphology of Sesotho as, additionally, they are agreement markers, also known as concords. These morphemes establish agreement between nouns, pronouns, qualificatives and verbs (see subsequent sections for these POS) and these concords are named according to the POS they link with the noun, for instance, subject-verb-agreement marker (SVA or SC) where SC is an abbreviation for subject concord, object agreement marker or OVA or OC for the subject verb and object verb agreement or adjective-agreement-marker etc. These morphemes are all categorized into classes that correspond with those of the noun.

Noun Class	**NUM**	**SUBJ**	**OBJ**	**ADJ**	**REL**	**ENUM**	**POSS**
1-2	SG PL	o- ba-	mo- ba-	e- ba-	ya- ba-	o- ba-	wa- ba-
1a-2a	SG PL	o- ba-	o- ba-	o- ba-	o- ba-	o- ba-	wa- ba-
3-4	SG PL	o- e-	o- e-	o- e-	o- e-	o- e-	ya- ya-
5-6	SG PL Extra PL	le- a- di-	le- a- di-	le- a- tse-	le- a- tse-	le- a di-	la- la- tsa-
7-8	SG PL	se- di-	se- di-	se- di-	se- di-	se- di-	sa- tsa
9-10	SG PL	e- di	e- di-	e(n)- tse(n)-	e- tse-	e- di-	ya- tsa-
14-6	SG PL	bo- a-	bo- a-	bo- a-	bo- a-	bo- a-	ba- a-
15	SG	ho-	ho-	ho-	ho-	ho-	ha-/hwa-
16-18	SG PL	bo-/ho- a-	bo-/ho a-	bo-/ho- a-	bo-/ho a--	bo-/ho- a-	bo-/ha-/hwa a-/ha-

Table 2.4: *Agreement/Concord markers of the noun and the qualificative in Sesotho*

Table 2.4. above also complied from Doke and Mofokeng (1957), illustrates the markers that link the noun to the verb and different sub-classes of the Qualificative and their complexity arising from the complex noun system. These markers prefix to the stem of the related sub-class to the noun it is linking to the qualificative. For instance, ***me****tse e* ***me****tle* ‘beautiful villages’ where class 3 noun prefix *me-* establishes agreement with the adjective stem *-tle* ‘beautiful’ through the prefix *me-*. The Subject (SUBJ) and object (OBJ) agreement markers will be

discussed below Section 2.2.2. under the verb. Though the same process applies to the pronoun and its sub-categories, it has not been included here due space restrictions but the different pronominal prefixes which abound in the corpus, will form part of the dictionary.

Sesotho nouns can also be divided according to the status of their origin into primary, derivatives/complex nouns, compounds, reduplicated and borrowed/loan nouns. Primitive nouns form by inflecting class prefixes to primary roots/stems such as *mo* + *tho* = *motho* 'person'. And a single stem can be inflected to different classes to increase the vocabulary stock as in -*sotho* which can attach to class 1 to form *mosotho* 'one whose allegiance is with the Basotho', class 7 *Sesotho* 'the language of the Basotho' class 14 bosotho 'the state of being a *mosotho* and class 5 *Lesotho* 'a country of the Basotho'.

Nouns also derive from other classes such as pronouns, adjectives and verbs to form complex nouns consisting of one root/stem and one or more affixes, for example:

i) Pronoun: *boyena* 'state of being him/her' < *yena* 'he/she'
ii) Adjective: *botle* 'state of being beautiful' < -*tle* 'beautiful'
iii) Verbs: *pina* 'song' or *mmini* singer' < bina 'sing' where the stem -*min*- is bounded by the prefix /m-/ and nominalising suffix /-i/

Compounding differs from derivatives in that it involves at least two roots/stems and one or more affixes in its formation. These formations, akin to primitive and complex words, are subjected to noun classification and concordial systems. A compound such as *moselakgomo* 'oxtail' arises from ***mo****sela 'tail'* + *kgomo* 'ox' where the prefix of the first member of the compound determines its class as class 3 and its plural as *meselakgomo* 'oxentails' and the prefix of the second member is dropped. Another type of compound is in the were a verb and noun combine to form a word like in the word *seilatsatsi* 'one who avoids the sun' made up of

the verb stem *-ila* ‘avoid’ and the noun *letsatsi* ‘sun’. The class 7 prefix *se-* is attached to *-ila* and the class prefix *le-* of letsatsi is done away with as per the rules of Sesotho.

Reduplication is yet another process by which Sesotho expands its vocabulary and changes the basic meaning of a word such as *motsemotse* < *motse* ‘village’ where the reduplicated form means the ‘real village’. How much of the stem is reduplicated depends on its length in terms of the number of syllables, for instance, *mosadi* ‘woman’ reduplicates into *mosadisadi*. The stem is disyllabic and so reduplicates in full whereas in *papadi* with a trisyllabic stem and an underspecified class 9 prefix */n-/*, it becomes *papadipadi* where only the last two syllables are the reduplicant. Where the stem is monosyllabic like in *motse* ‘village’, the stem together with the prefix will reduplicate and this is the only exception where a prefix is allowed to repeat itself as under general circumstances this is not allowed.

Loans/borrowed words or adoptives are an inevitable result of language contact and Sesotho displays a very good number of these and their adoption has impacted the Sesotho alphabet accordingly. Though they can be found among many classes, they tend to congregate in class 9. Most of them are phonologized into Sesotho and so follow the noun classification system like all other kinds of nouns, e.g., class 7-8 *sekete* ‘skirt’ becomes *dikete* ‘skirts’ and *hempe* ‘shirt’, *dihempe* ‘shirts’ is in class 9-10.

The noun is the category with the highest number of items (Section 4.2) as its function is to name entities. The corpus contains all these noun types, and at this stage, they are not singled out so that the focus can only be on them as POS. Other Sesotho POS include the qualificatives, which constitute adjectives, relatives, enumerative and the possessives. It has been argued that Sesotho does not have prepositions (Doke, 1957). The verb and pronoun, as well as qualificatives, are highly related to the noun. These classes substantiate the noun, and because of this, this information is important to note in lexicographic work because in this

language a word can belong to different classes and have different meanings as per classified usage.

2.2.2. Verb Structure

The main characteristic of the Sesotho verb is in the default tense. The present and other tenses except the past and the 3rd person imperative in both positive and negative form, and when it extends its meaning through suffixes, it ends in */-a/*. Syntactically, the verb governs the sentence, that is, the subject and the object where the latter occurs in a sentence. As such, it is not surprising that its relationship with these constituents is intricate. The verb always agrees with its subject in number and in Sesotho this agreement is shown by a compulsory member of the verb, the SVA or SC in order to establish the agreement between the two. The SVA arises from the class prefix of the subject noun and in a case where there is an understood object in the sentence, it will stem from that object, for example:

i) ***ba****na* ***ba*** *ja sehwapa* 'children are eating biltong.'
ii) ***mo****shanyana* ***o*** *bolaile tweba* 'the boy has killed a rat.'
iii) *moshanyana o* ***e*** *bolaile* 'the boy has killed it.'

Sentences i - iii above illustrate the relationship between the verb and the subject and/or object. In (i) the first bolded */ba-/* is a class 2 prefix of the stem */-na/* in *bana* 'children' and the second */ba-/* is the SVA linking the verb to the subject *'bana'.* The subject is in the plural and so the SVA is also in the plural. In (ii) there is an object *tweba* 'rat' where the verb *bolaile* 'killed' required an object to complete its meaning. The SVA /-o/ is singular. However, in (iii) which is the rendition of (ii) but with an understood object, /e-/, SVA class 9 stands for the understood object *tweba* and carries singularity like the object it stands for.

The SVA, like the default suffix /-a/ are two of the major characteristics of the verb. Since the verb is a word that denotes action, it also carries tense - the time when its action is happening- as its third main property.

Further, the Sesotho verbs use extensional morpheme to distinguish different expressions of a single verb. In such cases, the verb stem does not change but rather the extensions added onto it change depending on the how, when, and to whom the action is being applied. The extensional are tabled below:

Type Affix	**Ext**	**Extended V**	**Meaning**	**Derived Noun**	**Meaning**
Passive	-w-	abjwa/ajwa	be given	Aba	give
Applied	-el-	Abela	distribute to	Aba	give
Causative	-is-	Bisa	make to distribute	Aba	give
Reciprocal	-an-	Abelana	distribute in return	abela	give back
Intensive	-isis-	Abisisa	spread vigorously	Aba	give
Perfective	-ell-	Abela	distribute with	Aba	give
Extensive	-ak-	Abaka	busy distributing	Aba	give
Reversive	-oll-	Abolla	take back	Aba	give
Associative	-ahan-	N/A	N/A	N/A	N/A

Table 2.5: *Types of verbal extensions*

2.2.3. The Qualificatives

The Sesotho qualificative is divided into seven types as demonstrated below. It is a form that qualifies nouns or pronouns and the structure of a qualificative is determined by the prefix of the noun that precedes it. The substantive (noun or pronoun) qualified can be specified in a sentence or phrase or referred to anaphorically. Thetso (2018: 33) maintains that "The qualificative is the most important part of speech which gives extra/more information about the substantive in order to enhance both written and spoken communication."

Type	Examples	Construction
Adjective	*Motho* ***e mobe*** 'ugly person'	e + mobe
Demonstrative	*nku* ***yane*** 'that sheep'	y + ane
Enumerative	*basadi* ***ba babedi*** 'two women'	ba + babedi
Relative	*lejwe* ***le letle*** 'beautiful stone'	le + letle
Interrogative	*motho* ***mang****?* 'which person?'	ma + ng
Possessive	*ntho* ***ya ka*** 'my thing'	ya + ka
Quantifier	*bana* ***bohle*** 'all the children'	b + ohle

Table 2.6: *Types of qualificatives*

Substantives influence the form of a qualificative, particularly the stem. For example, in the case of the adjective in 2.6., ***e mobe***, if the Class 1 Singular Noun *motho* 'person' was to be replaced with Class 9 Singular Noun *ntja* 'dog', the phrase would change to *ntja* ***e mpe***. The concord **mo-** applies to Class 1 Nouns, which is the class for people. Whereas the concord **n-** falls under Class 9 Nouns, a class for animals. Therefore, the influence of **n-** causes modification in the stem ***mobe*** causing it to change to ***mpe*.** The Sesotho adjective agreement concord causes the link to the noun stem by prefixing a class marker to the stem.

When it comes to the demonstrative, as shown with ***yane*** in Table 2.6. the prefix ***y*** is agglutinated to the demonstrative stem **–ane**. The prefix in the demonstrative corresponds to the noun class prefix.

Below are examples that show the class prefixes' influence on the demonstrative prefix:

ngw*ana* 'child' (Singular, Class 1 Noun) ***ya****ne* 'that' (determiner)
ba*na* 'children' (Plural, Class 2 Noun) ***ba****ne* 'those' (plural determiner)

Although ***ngwana*** begins with a letter that is common to Class 9 nouns, the /*n*/ should not be read in isolation but as a single sound construction with /**gw**/ that has occurred because of the

process of velarisation. The etymological construction of the word *ngwana* is ***moana*** = ***mo*** (class 1 noun prefix) + ***ana*** (stem). The /m/ has velarised to /ng/ as a result of ***o*** + ***a*** forming a palatal glide /**w**/. However, the influence of the noun upon the demonstrative is clear in the plural demonstrative prefix, where both prefixes fall under Class 2.

Then there is the enumerative. Its relation to the substantive is seen in the concord and enumerate or number prefix as in Table 2.6. above, ***ba****sadi* 'women' (Plural, Class 2 Noun) ***ba*** (Plural, Class 1 Concord) ***ba****bedi* 'two' (Plural, Cardinal Number). Sesotho only has two enumerative concords, the singular and plural, and the plural applies to all numbers except for the number 1. For example:

> ***mo****sadi* 'woman' ***o*** (Singular, Class 1 Noun enumerative concord) ***mo****ng* 'one' (number)
>
> ***ba****sadi* (Plural, Class 2 Noun) ***ba*** (Plural, Class 1 Concord) bararo 'three', ***ba****ne* 'four', ***bahlano*** 'five' (Plurals, Cardinal Numbers)
>
> ***ba****sadi* ***ba tsh****elela* 'six', ***s****upa* 'seven', ***r****obedi* 'eight', ***r****obong* 'nine', ***l****eshome* 'ten'

The fourth qualificative is the relative, which on the one hand, can be a root and noun. In Table 2.6. there is ***lejwe*** 'stone' (Singular, Class 5 noun) ***le*** (concord) ***letle*** 'beautiful' (singular, Class 5 noun). The ***–tle*** **in** ***letle*** is also a root and can take various prefixes depending on the noun prefix. For example:

> ***mo****tho* ***e mo****tle*
>
> ***ma****tlo* ***a ma****tle*

The fifth is the interrogative. Interrogatives appear as question phrases and are marked by /**–ng/** or /**–fe/** as a suffixes. The example in Table 2.6. above ***ke mang*** 'who (person) is' is an example of anaphora in quantitative phrases and can also be phrased as ***ke ofe*** 'which (person)

is. In these examples, ***mang*** is a nominative/question whereas **o-** is a Class 1 Concord and ***–fe*** is a relative stem.

> *ke mang?* 'who is it?' < *motho eo le buang ka ena* ***ke mang****?* 'who is the person that you are talking about?'
>
> *ke ofe?* 'which one?' < *motho eo le buang ka ena* ***ke ofe****?* 'which one are you talking about?*'*

The sixth quantitative is the possessive, which is regarded as a determiner. In terms of structure, the possessive concord is determined by the class of the subject concord.

In addition to the example in Table 2.6. above, here are other cases that show this:

> ***ntlo ya ka*** 'my house'
> ***n****tlo* 'house' (Singular, Class 9 noun) ***ya*** (Class 9 possessive concord) ***ka*** (1st person possessive stem)
>
> ***lemati la ntlo*** 'the house door'
> ***le****mati* 'door' (Singular, Class 5 noun) ***la*** (Class 5 possessive concord) *ntlo* (Singular, Class 9 noun)
>
> ***mmino wa setjhaba*** 'the nation's music'
> ***m****mino* 'music' (Singular, Class 3 noun) ***wa*** (Class 3 possessive concord) *setjhaba* 'nation' (singular, Class 7 noun)

Lastly, the quantifier. Sesotho has one fixed quantifier, ***-hle***. It takes the form of a pronoun and operates slightly similar to the adjective because the quantitative pronoun is agglutinated to any prefix it takes, and the prefix is determined by the prefix of the noun that the quantitative modifies as seen in Table 2.6., ***ba****na* 'children' (Plural, Class 1 noun) ***bo****hle* (Plural, Class 1 quantitative pronoun). Other examples are:

> **ma**mati 'doors' (Plural, Class 6 Noun) **o**hle 'all' (Class 6 Quantitative Pronoun)

dibuka ‘books’ (Plural, Class 10 Noun) **tso**hle ‘all’ (Class 10 Quantitative Pronoun)

2.2.4. Morphology of the Adverb

The adverb is one of the open word classes and its function is to describe the verb or the qualificative or another verb (Doke and Mofokeng: 1974). There are three categories of the adverb, which are time, manner and location. Some adverbs can also function as Classes 16-18 nouns; however, the adverb is not linked to the agreement/concordial system of the noun. In some instances, the adverb is preceded by adverbial concords. Below are examples of adverbs and a description of how they are formed from other parts of speech.

<u>Time</u>

i) *ntate o tsamaya hoseng* “father leaves in the morning”

ntate (subject) *o* (Class 1 SVA) *tsamaya* (Verb) *hoseng* (Adverb)

<u>Manner</u>

ii) *ntate o tsamaya ha butle* “father walks slowly”

ntate (subject) *o* (Class 1 SVA) *tsamaya* (Verb) *butle* (Adverb)

<u>Location</u>

iii) *ntate o tsamaya tseleng* “father is walking on the road”
iv) *ntate* (Subject) *o* (Class 1 SVA) *tsamaya* (Verb) *tseleng* (Adverb)

Finally, to determine whether there is an adverb in a phrase, one must look at whether the word that follows a verb describes *when* (time), *how* (manner), or *where* (location) the action is taking place. For example: In (i) the adverb *hoseng* ‘morning’ describes the verb *tsamaya* ‘leave’ in that it specifies the time *ntate* will leave. Unlike in (ii), the adverb in this sentence is *butle ‘slowly’* and it is preceded by adverbial agreement marker *ha*. In (iii), the verb is qualified by the adverb *tseleng* ‘on the road’ and *tseleng* is an adverb derived from the noun *tsela* ‘road’. Sesotho locatives are identified by suffixing *-ng* onto nouns.

2.2.5. The Pronoun

Pronouns are used in the place of nouns. Four types of pronouns have been identified in Sesotho (Olivier, 2009). These are the absolute, demonstrative, qualitative and qualificative pronouns. Pronouns can appear individually without the use of agreement concords. The relationship between the pronoun and the noun it refers to or represents can be seen in the prefixes.

First looking at the absolute pronoun, the stem *-na* is constant for absolute pronouns even when the class prefix changes in line with the class of the noun represented, and these pronouns translate to 'she/he', 'it', 'they' or 'them':

i) Motho "person" – e**na**: *ena (motho) a buang* "she is the one talking"
ii) Seeta "shoe" – so**na**: *sona (seeta) sa hao* "it is yours"
iii) letsatsi "day" – lo**na**: *lona (letsatsi) le fihla* "when it arrives"

In the case of the demonstrative pronoun, it has three positions. The 1st position demonstrative pronoun refers to the agent that is in the environment. They are formed with a class prefix and a stem. The bolded words below all translate to 'this':

i) (motho) **enwa**
ii) (seeta) **sena**
iii) (letsatsi) **lena**

The 2nd demonstrative refers to the agent in relative distance to the speaker, which translates to either the singular 'that' or plural 'those'.

i) motho **eo**
ii) seeta **seo**
iii) letsatsi **leo**

The 3rd demonstrative pronoun refers to a third person or object away from the speaker and the agent spoken to. The bolded words mean that the person or things referred to 'over there' or 'yonder'.

i) motho **yane**
ii) seeta **sane**
iii) letsatsi **lane**

Qualitative and quantitative pronouns function in their primary form. In Sesotho, they describe the noun more than they refer to it. Sesotho pronouns are used in the same sentence as the noun they are describing.

The bolded word below plays the role of a quantitative pronoun:

i) *letsheare* ***lohle*** 'all day'

The ***-hle*** in *lohle* is classified as a quantitative pronoun stem, and ***-hle*** it is also a quantitative stem. In this case, *lohle* plays the role of describing the noun *letsheare* 'day'and *letsheare* could well be replaced with a Class 5 absolute pronoun *lona* and the phrase would read as *lona lohle* "the whole of it".

It is the same with the Sesotho qualificative pronoun. This pronoun is a relative - a type of qualificative - and it is used to describe the noun. For example:

i) *lekgarebe le lebe ke leo* "here comes the ugly girl"

To demonstrate how the relative phrase *le lebe* qualifies as a pronoun, the sentence can stand alone without the noun *lekgarebe*:

i) *le lebe ke leo* "here comes the ugly one"

Different types of qualificatives are discussed below where the quantitative and relative will be discussed further.

2.2.6. Fixed Categories

The minor classes, which are also defined as functional classes as shown in Table 2.2. under closed classes, include conjunctives, clitics, and interjections and as a general rule they do not inflect, and each is discussed briefly below:

Conjunctives

Conjunctives are words whose function is to link up words or phrases or sentences or serve as introduction and these are called conjunctives. Though they are also radical in nature, they do sometimes occur as inflected forms. Examples of radicals are *kapa* 'or', *mme* 'but', *athe* 'whereas' and those of inflected forms are *jwale* 'therefore', *ho fihlela* 'until', *le jwale* 'and indeed'.

Interjections

Interjections are a group of words of exclamatory nature and are used to express emotion or call for attention or give commands or indicate agreement or dissent (Doke and Mofokeng 1957:365). They are roots and do not take part in the concordial system of the noun. Examples thereof are *aha!* 'joyous vindication' *awo!* 'expression of sympathy,' *be!* 'surprise' and so on. Some pronouns and nouns can be used for exclamatory effect and will take an exclamation mark to indicate their changed status, for example, *banna!* 'man!' or *wena!* 'hey!'.

Sesotho also has a few clitics which are morphemes that are not clearly independent words or affixes. Sometimes they are known as "leaning-on words" Spencer (1998). They are open monosyllabic morphemes such as *-yo*, 'not there', *he-* 'then' hle 'please' etc. as in the following examples:

(i) *banna ha ba* ***yo*** 'the men are not there'

(ii) *binang* ***he*** 'sing then'

(iii) ipateng, ***hle*** 'please hide yourselves'

2.2.7. Hybrid classes

Generally, ideophones and pronouns are classified under fixed classes but a closer study in Sesotho has shown that both can inflect. Ideophones can derive from verb stems with a resultant category change which in fact is a main consideration in the formation of complex words according to Wardhaugh (1972) in their discussion of differences between derivation and inflection. Pronouns can be inflected to form nouns, as in *boyena* 'herself/himself'. Their inflection changes category and Doke and Mofokeng (1957:338) attest to the fact that they do inflect. Each is discussed briefly below.

Doke and Mofokeng (1957: 337) describe ideophones as onomatopoeic words whose function is similar to that of adverbs in that they qualify or describe the verb as to manner, colour, sound, smell, action state and intensity. Their roots are said to be irreducible, that is, they cannot be broken down any further. They occur in constructions following the verb *ho re* 'to be' as in the example (i) or as a qualifier in (ii) below. In (i) *twa* 'blindingly white' is the ideophone and *twa* is the radical form. In example (ii) *tlere* 'very red' describes the intensity of the colour *kgubedu* 'red' and so functions as an adverb describing the adjective *kgubedu.*

(i) *ho re* construction: *ho re twa kantle.* 'It is snow white outside'

(ii) Descriptor construction: *Pule o mo kgubedu tlere.* 'Pule is very red'

(iii) Reduplication: *ngwana o bata po! po! po!* 'the child is extremely cold'

(iv) Derivation: *banana ba ba rata rati!* 'The girls love them intensely'

They are also known to reduplicate extensively as in (iii) where *po* 'very cold' has reduplicated to *po! po! po!* to heighten intensity. Finally, they also derive from verbs as in (iv) where the verb stem *rata* 'love' becomes *rati* as an ideophone and qualifies *rata.*

2.3 Corpus Linguistics in lexicography

A corpus is a collection of naturally occurring language texts that serve as a basis for linguistic analysis or description (Sinclair, 1991: 168). It is textual proof of the structure and use of a language, made up of samples of texts in written and spoken form (Kennedy, 1998). The advantage of a corpus as a tool for language study is that its results are quantifiable and comparative for another research (Biber et al., 1998). Kennedy (1998) brings to attention that corpora building has a different purpose, and the purpose determines the size, design, and nature of the corpus. The dictionary initiated in this research comprises of words extracted from Sesotho written texts and because of the objective to serve as a reference tool for school learners, these texts had to include activity books, examination papers, and literary texts used in schools that teach Sesotho. Moreover, because the dictionary is a pilot to a larger dictionary transaction, the size of the corpus is small in comparison to the initial conventional number of standard corpora, which is one million words.

What gives the Corpus Linguistics approaches approbation is the use of both quantitative and qualitative methods for analysis. Corpus Linguistics is one source of evidence for improving descriptions of the structure and use of language and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language. A corpus designed for linguistic analysis usually is a systematic, planned and structured compilation of texts (Kennedy, 1998). A corpus constitutes an empirical basis not for identifying the elements and structural patterns which make up the system used in a language, but also for mapping out the use of these systems (Leech, 1991:11). Projects such as

the British National Corpus (1994), Oxford English Dictionary (Kennedy, 1998: 11), and Collins Birmingham University International Language Database (COBUILD) (1980) are some of the exemplary projects that adapted the Corpus Linguistics approaches to dictionary-making, and these projects continue to expand both their corpora and continue to create different lexicography products.

Some Southern Africa languages have adapted the Corpus Linguistics methodology to lexicographic practice in recent years; one such language is Setswana. Otlogetswe (2013) on the development of a monolingual Setswana dictionary, a sister language of Sesotho, which is used in both Botswana and South Africa highlights that Setswana, like Sesotho, has also experienced low dictionary production over the years. Otlogetswe (2013) provides insight to how Corpus Linguistics is used towards the development of African languages dictionaries; this briefly corroborates the usefulness of this methodology for resource-scarce languages in the age of language computing.

Leech (1991:11) asserts that for a corpus to be representative of the different varieties of a language, it should comprise of samples from different texts in a language. Sinclair (1985: 81) maintains that "The main sources of lexicographic evidence are, probably in the order of popularity: 1. other dictionaries, 2. introspection, and 3. observation of language in use." Sinclair (1985) advises that one apply partial reliance on existing lexicography when developing a new resource and maintains the following regarding existing dictionaries,

(a) It is difficult to know when a word or a usage lapse; hence for a contemporary record the evidence of existing dictionaries is misleading.
(b) Structural mistakes in an entry are difficult to detect since they only become noticeable when compiling starts from scratch.

(c) New developments in linguistic description, for example, pragmatic information, cannot easily be fitted into established lexicographic formats or indeed into the whole conceptual background of lexicography. The received information cannot be shaken free of its theoretical origins.

The above are listed as disadvantages, and in the case of Sesotho, the existing bilingual and trilingual dictionaries do not indicate the usage of words. Even though one were to rely on personal knowledge, this would not reflect "various kinds of language patterning psychological associations, semantic groupings, etc." As a result, Sinclair (1985) guarantees that a corpus is an efficient tool in lexicographic practice. "The selection process becomes a selection of texts for the corpus, not instances for a dictionary. Once it is decided to include text, then all the instances of all the words constitute the evidence" (Sinclair, 1985: 83).

Further, Sinclair (1985: 83 – 84) asserts that,

> The technique of creating a corpus for the linguistic study is becoming more and more important as the computers become capable of processing longer and longer texts, and of increasing the sophistication of the processing […] More specialized corpora on various principles are in various stages of planning and creation and one can confidently predict a period of great activity in this field, fuelling the lexicography of the nineties and beyond.

A number of lexicography projects for African Languages have used corpora as tools for harvesting lexical data (Otlogetswe, 2003), and have reported effective results.

Meta-lexicography, as outlined by Wiegand (1983), is useful for understanding the relationship between lexicography and Corpus Linguistics. For example, Wiegand (1983: 15) illustrates how the different components of general lexicography merged with other theories to study the language. These components highlight some of the aspects for consideration such as the purpose of dictionaries in general society, which gives rise to the question concerning the uses of a monolingual dictionary for L1 speakers of the same language in a multilingual society.

According to Wiegand (1983: 17), "When lexicographers explain the meaning of a lemma-sign in a monolingual dictionary […] they must in principle proceed the same way speakers of the language do in everyday dialogues about the meanings of parts of utterances that have been questioned." This statement supports and somehow credits the method used for data gathering in this research, whereby the assessment and evaluation of word meanings came from a multi-genre corpus, in addition to examining Sesotho Bilingual dictionaries.

Kytö and Rissanen (1993:5) outline four aspects of consideration in building a diachronic corpus. The corpus should have chronological, regional, sociolinguistic, and generic coverage. The corpus in question covers all these. The texts that make up the corpus compiled for this research are in South African and Lesotho Sesotho, this covers the regional aspect. On sociolinguistic coverage, which mainly looks at gender, both female and male written texts are represented. However, this is not on a fifty-fifty basis because there are not a lot of female authors in Sesotho literature.

On the aspect of chronology, the hymns, for example, have existed since the early period (1800) of reducing Sesotho into writing, and there are those texts that were produced in the 21st century such as the short stories, for instance. Another aspect, which is coverage, is also accounted for in the compiled Sesotho corpus. Generic coverage is crucial because it relates to the issue of representativeness, which is a crucial factor in corpus design (Biber et al., 1998). Even though the genres do not contain the same number of texts, each genre has at least 1000 words**.** The COBUILD dictionary project, set up by Collins publishers and Birmingham University in 1980, was the first to use a computer corpus as its primary basis (Sinclair, 1987). Even so, Krishnamurthy (2006: 4) asserts that "For native-speaker dictionaries, the largest corpora currently available (over 500 million words) still cannot supply sufficient evidence to distinguish the normal use of rare words and phrases from chance and arbitrary usage." This demonstrates the challenges that arise with the changes that occur in a

language over time; however, it also makes it important to have records of a language every day. Krishnamurthy (2006: 9 – 10) states that,

> Dictionaries list the words that occur in the texts of a language, and should, therefore, contain generalizations about the word usages in those texts. However, few dictionaries in the past have supplied the textual evidence their entries were based on, so users have had to rely on the ability and probity of the lexicographer.

Corpus Linguistics is used in this research as a methodology to study Sesotho usage in social contexts by studying its structure from different text genres for the further description and explanation of its nature in both spoken and written form. Texts are the source of evidence for linguistic description and study. When the texts of a language are compiled together to study the language, they are called a corpus, which represents a natural language and reflects its state or variety (Sinclair, 1991: 2). Leech (1992: 107) highlights that the focus of study in Corpus Linguistics is the performance rather than competence, and observation of language in use leading to theory (Kennedy, 1998: 7). Arguably, Corpus Linguistics is not a theory or branch of linguistics but a methodology to study aspects of linguistics.

Although it may not be a branch of linguistics, Corpus Linguistics allows research to make distinctions between methodological approaches taken to linguistic areas of enquiry. There was no precise definition of corpus practices during its initial use for language study in 1813 up to the 1950s, the early period of Corpus Linguistics. Boas (1940) used the corpus-based approach, then referred to as field linguistics because of how texts/utterances were collected, which was an approach of corpus-based description. The early practice of Corpus Linguistics saw a decline following the criticism offered by Chomsky in the 1950s. Chomsky suggested a rationalist from the empirical view of linguistics, which disapproves the use of a corpus as a tool for language study given the role of the linguist as modelling language competence rather than performance (Fort, 2011).

Following the criticism that Corpus Linguistics faced between the mid-1950s to the 1970s, the methodology was resuscitated in the 1980s and continues to be used to date with credible results to language study (McEnery and Wilson, 1996: 4). Some of the developments in the field of Corpus Linguistics over the years include microcomputers from the 1970s and the advent of the CD ROM in the 1980s. These made Corpus Linguistics research more accessible to a much fuller range of participants. Most electronic databases were compiled and made available in the 1990s, for various purposes (Kennedy 1998, 4). Some of the most prominent corpus projects have been undertaken for commercial purposes by dictionary publishers, and this is what this research anticipates achieving for Sesotho as well.

Kennedy (1998: 2) maintains that "Corpus linguistics did not begin with the development of computers, but there is no doubt that computers have given corpus linguistics a huge boost by reducing much of the drudgery of text-based linguistic description and vastly increasing the size of the database used for analysis." Linguists use corpora to answer questions and solve problems. Kennedy (1998) further asserts that the most critical skill is to be able to ask insightful questions that address real issues in theoretical, descriptive, and applied language studies. This emphasises the point that corpora are built with purposes, and it is the purpose that will determine the size, design and nature of the corpus (Kennedy, 1998).

Based on the above, a corpus does not have to be a mere body of texts that describe both the structure and use of language, (Kennedy 1998, Biber et al. 1998). It requires proper organisation and should meet specific standards to be useful for language research. It has to be a maximally representative finite sample that will enable results to be quantified and compared to others in the same way as any other scientific investigation, which is based on a data sample (McEnery and Wilson, 2001). Atkins and Rundell (2008: 75) highlight that, "A truly representative corpus is an impossible goal because we are sampling from a population whose nature is unknowable and whose extent is unlimited."

The availability of digital computers from the 1980s brought about a radical change in text-based scholarship. During this period, lexicographers also used computers to store large amounts of text examples and retrieve particular words, phrases or whole chunks of texts in context (concordance tools), quickly and exhaustively, on their screens. The availability of digital computers make corpus processing efficient, with Natural Languages Processing tools that are accountable based on the language's development of such tools, computers provide accurate replicability, statistical reliability and they can handle large amounts of data.

Fillmore et al. (1988) add that there are flaws that can arise in building a lexicon if the source, a corpus, for example, becomes "unbalanced". Therefore, to avoid this, the compiler should strive for the maximum representation of the language genres in their sampling of corpus data. Also, Otlogetswe (2013: 536) argues that "if the corpus data is skewed and badly structured, the data abstracted from the corpus will also reflect this - one cannot take from a corpus what has not been put into it."

Biber et al. (1998) have depicted the usefulness of a corpus-based approach to lexicography. Their focus is on the analysis of words from a lexicon, even more, their account is relevant here because it illustrates the research question that can be addressed using a corpus regarding the meanings associated with single words, and a defining dictionary would be required for addressing this question. From this standpoint, the corpus-based approach in lexicography comes forth as a method for solving other language queries such as ambiguity. Lexicography requires the extensive use of corpora because word senses and collocation patterns are often much less common than grammatical patterns and a corpus has the properties to assist language researchers in understanding the meanings of words as well as their distribution and use. It is evident that there are different ways of using a corpus as a research and application tool in lexicography. Some advantages of the corpus tool in lexicography

include the ability to identify synonyms and the fluidity to extract and layout their usage examples in a defining dictionary.

2.3.1. Lexicography

Lexicography involves the compilation of dictionaries; it is the compilation of dictionaries based on theoretical conventions that inform the process. For example, taking from Bergenholtz et al. (1997 in Bergenholtz and Gouws, 2012: 38), "There are two types of lexicography: [first] the development of theories about and the conceptualisation of dictionaries, specifically concerning the function, the structure and the contents of dictionaries. This part of lexicography is known as meta-lexicography or theoretical lexicography." This type of lexicography lays out the guidelines to carry out the activity of dictionary-making. The second part is "the planning and compilation of concrete dictionaries. This part of lexicography is known as practical lexicography or the lexicographic practice," Bergenholtz et al. (1997 in Bergenholtz and Gouws, 2012: 38). With these distinctions, it is understood that one cannot carry out lexicographic practice without orienting oneself with meta-lexicography, at least for efficient practice.

Lexical dictionaries for languages of the world started to be compiled in 32nd century BC with the earliest recorded dictionaries said to be the monolingual Sumerian wordlist in cuneiform writing (Durkin, ed. 2016). In 24th century BC, the initial bilingual wordlist of Sumerian and Semitic language Eblaite was compiled. Further down the years, in the 18th century, the thematically ordered Egyptian list of nouns was compiled. The tradition has continued since with modern languages starting in the same direction by developing wordlists, some of which were and continue to be expanded into dictionaries.

From observing the different dictionaries it is noted that dictionaries have some standard aspects to them; however, Bergenholtz and Gouws (2012: 36) state that when it comes

to meta-lexicography, “there are different lexicographic theories”. Theories may differ depending on the language type and depend on the time at which the lexicographic practice occurs. For example, over the years, there have been different theories concerning where to position lexicography in academia, and over time the real aspect of meta-lexicography surfaced. Bergenholtz and Gouws (2012: 36) add regarding lexicography theories that “[…] they all acknowledge the fact that the lexicographic practice is complemented by a theoretical component and that lexicography, with dictionaries as its subject matter, should be regarded as an independent discipline.”

In language studies, some scholars view the theory and practice of lexicography as a topic of applied linguistics (Hartmann, 1996). Hartmann (1996) maintains that the criteria for a discipline of applied linguistics are that “it should be linguistic in orientation, interdisciplinary in outlook and problem-solving in spirit.” McArthur (1998) and Bergenholtz and Gouws (2012), on the other hand, view lexicography as a reference-science. Bergenholtz and Gouws (2012: 76) hold that lexicography is a part of information science and not a sub-discipline of linguistics. Howard (2002) on another hand asserts that lexicographic practice has always drawn on a range of other disciplines and crafts, relevant to reference works generally, it is widely considered as either discipline in its own right (sui generis) or a branch of reference science or information science.

The differences in the categorisation of lexicography are concerned about information science rather than emphasising the storage aspect of lexicographic information - overlooking that lexicography develops from the scientific study of language and social language needs. There needs to be an inclusion of the compilation of lexicons as reference tools from which information is retrievable. It is, therefore, reasonable to look at lexicography as a discipline in its own right, when bringing in the aspects of reference. Lexicography qualifies as information science as well as it does to applied linguistics when considering that it is a method of

preserving the language, that indicates word classes and from which language use and structure can be studied in both print and electronic media.

Wiegand (1983), however, refutes the assertion that lexicography falls under applied linguistics, and concurs that it is a form of science. Wiegand (1983: 14) writes on a general theory of lexicography and recognises lexicography as neither a science, nor a branch of applied linguistics, or a branch of lexicology and that its production forms part of reference work. Additionally, linguistic lexicography is categorised as "scientific practice aimed at producing reference work for a language" (Wiegand, 1983: 14). Based on this rationale, Mother Language (L1) dictionaries are an essential asset for speakers to keep up to date with the new concepts that are introduced into their cultures and for linguistic preservation alongside linguistic reference.

This study, therefore, agrees with Bergenholtz and Gouws (2012) on the conclusion that lexicography is a discipline on its own because, for example, the terminology development activity which Bergenholtz and Gouws assert as synonymous with lexicography, involves consultation with experts in a particular field. Bergenholtz and Gouws (2012: 40) add that "For the time being we, therefore, regard lexicography as an independent discipline, relying on experts coming from different disciplines." Therefore, it is accurate to say that lexicography is a field on its own that can be applied to other disciplines. "We regard the lexicographic theory as a discipline not only directed at the production of dictionaries but in a more general way as the production of information tools" (Bergenholtz and Gouws, 2012: 40).

All the disciplines that support lexicography add to its wholesome implication with the components that stipulate the processes, tools and methods involved in both the practice and theory of lexicography. From the methods of information or reference science, lexicography conveys data using a semasiological approach. Warbuton (2015: 374) states that "This is an approach in which the linguistic form is the object of study." On the other hand, Rey (1995:

116) maintains that lexicography concerns itself with the function and the behaviour of words in society. It proceeds semasiologically, that is, it takes signs and studies their meaning, and ultimately, the concepts for which they stand.

As already established, the term lexicography has two implications. First, it refers to the compilation of dictionaries using various practical approaches. Secondly, it relates to the study of dictionaries (Jackson, 2002). This research looks at the scope of lexicography from both perspectives. McEnery (1992: 25) defines lexicography as the study of occurring word tokens, falling under the morphological aspect of linguistics, which allows for analysis of semantic and grammatical class membership. To corroborate this, Wierzbicka (1985) maintains that semantics must provide theoretical foundations for lexicography. This is intuitive as dictionaries are information tools and given that, semantics comprises of elements of sense, reference, implication, and logical form. These aspects are a part of single word tokens in dictionary entries. L1 speakers and learners of Sesotho can further advance their knowledge of the language from the monolingual dictionary; because it will provide meaning and parts-of-speech as well as usage examples, properties that most Sesotho dictionaries do not have. Rey (1995: 109) states that the need for compiling dictionaries became socially significant when the need to relate different languages became imperative, especially during the seventeenth century. At that time, these dictionaries were crucial in Spain, Italy, and France with English becoming the self- regulatory description of diverse languages.

Lexicography is also defined as a pragmatic than a theoretical concept. Its practices have to do with the development of jargon for a specific scientific field. In contrast, terminologies are groups of terms or nomenclature. Nomenclature is the totality of names expressed by nouns. In this way, 'terminology' would be reserved for the activities performed in relation to nomenclatures (Rey, 1995: 127). The domain of naming and designing, i.e. the various semantic processes by means of which the observed or

conceived ‘realia’ are attached to linguistic signs, and thus to units of the lexicon, after the abstraction enabling them to be matched to the symbols or a code.

Means that are used for term formation include having to synthetically-modify lexical items by means of affixes (Sager, 1990: 61). Synthetic-modification can be seen from a lot of borrowed Sesotho words, but this does not give lexical items much linguistic information, for example, the word *litheresi* for 'literacy' cannot be decomposed into morphemes and for semantic analyses because an ordinary Sesotho speaker or learner would not be able to identify the root of this word following the linguistic rules of Sesotho. The affixes are lexically meaningful items which in Sesotho cannot stand on their own. Moreover, because of this, synthetically modifying lexical items for term formation cannot completely be relied upon.

Another way is the analytic formation, which combines lexical units into larger units, i.e. compounding. Compounding is also found used instead of synthetic-modification in Sesotho, for example, the existing Sesotho online dictionary *Bukantswe* is made up of two words *buka* ‘book’ and *ntswe* being derived from *lentswe* ‘word, in this case. Modern terminological theory accepts the occurrence of synonymic expressions and variants of terms and rejects the narrowly prescriptive attitude of the past, which associated one concept with only one term. With this knowledge of terminology, it is going to be easy to understand foreign terms within the corpus compiled for this project. This theory will be helpful when making suggestions on such words - to understand the processes involved and substantially conceptualise meanings for the dictionary entries.

Dictionary research is divided into four components, i.e. history, typology, criticism and use, whereas dictionary-making is divided into three components, i.e. fieldwork, description and presentation. Burkhanov (1998) asserts that “lexicography has been successfully developing its theory. He also argues that “The term ‘lexicography’ refers to the

process, result, and theoretical evaluation, of the making of reference works which represent a wide range of heterogeneous knowledge structures."

Jackson (2002: 22) asserts that "We rely on dictionaries to provide us with definitions of words and tell us how to spell them […]. A dictionary is a reference book about words." A dictionary has a headword list and dictionaries are commonly arranged in alphabetic order, and the same is applied in this research (Chapter 4, and Appendix 1). "A dictionary is more than a reference book; it is also a (partial) record of the vocabulary of a language" (Jackson, 2002: 23). Jackson (2002: 23) states that dictionaries are used for various reasons by different people and "Sometimes we want to check the spelling of a word […]." A dictionary has a recording function and a reference function. For this research, the dictionary describes Sesotho vocabulary up to the current times and has sampled texts from different decades, i.e. the corpus is diachronic, and so is the dictionary.

2.3.2. Bantu languages lexicography

When it comes to African Languages lexicography, the first dictionary that was compiled is recorded as the *Vocabularium, Latinum, Hispanicum, e Congense* in 1652 by Van Gheel (De Kind et al., 2012). "Given Van Gheel's manuscript survives to this day, it is possible and even necessary to move the origin of the field back to 1652 or, writing in 2012, to state that the field of Bantu lexicography is (at least) 360 years old" (De Kind et al., 2012: 161). On the other hand, the African Association for Lexicography (AFRILEX) has existed since 1995, yet there is little said about Sesotho dictionary research in the AFRILEX research journal, *Lexikos* that was established in 1991. The metalexicography of Sesotho is far behind its practical lexicography. Additionally, it has become tradition that Sesotho dictionaries are bilingual and translate from English or/and Afrikaans to Sesotho or vice versa. Prinsloo and Mashamaite (1999: 233) make a critical point regarding the lexicographic tradition of African languages that,

> The history of lexicography and lexicographic research in African languages is marked and dominated by research output from non-mother-tongue speakers of the languages. Common examples can be found in the numerous dictionaries in each African language today. All compilers are non-mother-tongue speakers, and in some cases mother-tongue speakers are co-compilers. These mother-tongue co-compilers did not have any knowledge of the principles of lexicography and metalexicography in particular. Non-mother-tongue speakers fulfilled an important function in the development of the African languages. [However], it is important that the mother-tongue speakers now double their efforts and commitment to make it their task and obligation to write and compile dictionaries in their respective languages.

2.3.3. Sesotho Lexicography

In 1841 Eugéne Casalis authored a book recording Sesotho grammar (Mabille and Dieterlen, 1974: iii). The structure of Sesotho writing in Casalis' work takes from the Latin alphabet; this alphabet was further expanded with an addition of digraphs and trigraphs to accommodate the complex sounds in Sesotho that do not exist in Latin such as [ts] and [tlh]. However, Casalis (1841) did not represent the morphological and phonological aspects of Sesotho efficiently. As a result, Casalis later published the English – Sesuto Vocabulary in 1894 (Moleleki, 1999: 244). Nevertheless, Moleleki states that "True to its name, the work is nothing else but an alphabetically arranged list of English words with their Sesotho counterpart" (Moleleki, 1999: 244), which is equally a glossary because it does not provide the semantic properties of the words. This research acknowledges the Department of Education, Department of Higher Education and the PanSALB's current efforts to elevate the status of previously marginalised languages like Sesotho as the medium of instruction and not only a school subject. However, there is a more significant need for resources to support effective implementation. Thus, there is a need for monolingual dictionaries that will account for the current language learning and teaching needs.

At the time that Moleleki (1999: 247) wrote in his account regarding Sesotho lexicography, he observed that:

> From the […] overview of the state of lexicography in Sesotho, it becomes clear that Sesotho is in dire need of a monolingual dictionary; a dictionary which will affirm the status of Sesotho as a language [while] contributing towards the formation of a healthy mindset with the Sesotho personality. However, such a project should be seen as a long-term undertaking because of its magnitude and complexity.

In as much as non-mother language speakers of African languages created the lexicography path for languages such as Sesotho, most of this work; speaking for Sesotho and the case of Sesotho bilingual dictionaries, is not comprehensive. The improvement is seen, however, in a bilingual dictionary that was edited by Motsapi (2015) under the *Sesiu sa Sesotho National Lexicography Unit.* What is surfacing from this observation is that a mother language speaker of a language is a better resource in compiling a dictionary for that language, but they would also have to be well versed in the target language. Further, as Prinsloo and Mashamaite (1999) mention, more than being a mother language speaker of a language, the core insight of lexicography conventions is equally important.

What is said above for lexicographers is in line with the argument of this study that, Sesotho mother language learners need Sesotho learning resources for them to know how to use the language whether in the classroom for spelling and definitions or to learn other languages adequately and efficiently. One could argue that if a learner cannot explain a mathematical equation in his first language, then they do not thoroughly understand mathematics. It is, therefore, vital to develop monolingual dictionaries and terminology lists that will support the understanding of education in learners' native languages, thus the initiative of this study to work towards the development of a monolingual dictionary for school learners.

It is also an established fact that for any lexicographic development to take root, the *sine qua non* is a vocabulary list and most importantly, a grammar book of that language, the two being two sides of the same coin. The latter is crucial in providing the semantic and syntactic information about any given word in the language. It is the foundation of the microstructure of any dictionary and can also determine how detailed the said information can be. As such, Casalis' grammar development efforts for Sesotho provided a fertile ground for the advancement of Sesotho lexicography. Though the first Sesotho dictionary was produced in 1859 from Rev. Mabille's vocabulary list on a long ship journey to Cape Town, it was only published in 1876 according to Ambrose (2006:20) in Motjope-Mokhali et al. (2020:261) and not 1878 as some literature has recorded. This made it the second dictionary in the South-Eastern Bantu family to be developed following John Brown's Setswana English Dictionary in 1875. In similar fashion, it was the work of a missionary, and they should be applauded not only for standardising Sesotho but also their footprint in both lexicographic and syntacto-semantic records of the language.

A survey of available Sesotho dictionaries follows below, and it is divided into two tables – Figures 2.7(a) and 2.7(b) - for purposes of readability and page organisation. The former table covers the period 1876-1965 and the latter 1965 to date. During 1876-1965, the individual lexicographic work of Rev. Adolph Mabille (and later Dieterlen) dominated the period, and the initial dictionary was expanded and went through eight editions.

Dictionary name	Date of Publication	Edition No./ Revision	Author	Type and size
Sesuto-English Vocabulary	**1876**	1st Edition	Adolph Mabille	Bilingual 158 pp. one column format
Sesuto-English Vocabulary	1893	2nd Edition	Adolph Mabille	Bilingual 144 pp. two column format
Sesuto-English Vocabulary	1904	3rd Edition	Adolph Mabille	Bilingual 267 pp. two column format
English-Sesuto-Vocabulary	1905	Only Edition	A. Casalis	Bilingual English Lemma list with Basic Sesotho Translations
English-Se-suto Vocabulary	Unknown	Only Edition	D.F. Ellenburger	Bilingual
Sesuto-English Dictionary	1911	4th Edition	Mabille and Dieterlen	Bilingual 511 pp. two column format
Sesuto-English Dictionary	1917	5th Edition	Mabille and Dieterlen	Bilingual 535 pp. two column format
Sesuto-English Dictionary	1937	6th Edition	Mabille and Dieterlen	Bilingual 495 pp. two column format
Sesuto-English Dictionary	1950	7th Edition	Mabille and Dieterlen	Bilingual 445 pp. two column format
Sesuto-English Dictionary	1961	8th Edition	Mabille and Dieterlen	Bilingual unspecified pp. two-column format
English-Sotho, Sotho-English Pocket Dictionary	1960	Only Edition	S. Christeller	Bilingual 144 pp. Bi-directional
English-Southern Sotho Dictionary	1965	Only Edition	L. Hamel (OMI)	Bilingual 8 Volumes in one column format
Se-Sotho-English Vocabulary	Unknown	Only Edition	T. Verdier	Unknown

Table 2.7a: *Timeline and Survey table of Sesotho Dictionaries 1876-1965*

At first it was named *Sesuto-English Vocabulary* and kept this name throughout the first three editions but changed its title to *Sesuto-English Dictionary* from the 4th edition onwards when it had been expanded in terms of number of entries and coverage of subject areas. All these editions are in the orthography of Lesotho though later, the same dictionary was published in the then new South African Republic orthography. The timeline from one edition to the next was 14.8 years except for the period 1904 to 1917 where it was 13 years.

There were other attempts by other missionaries such as Ellenburger, Hamel and Casalis but Mabille's has remained the mainstay of Sesotho lexicographic tradition. Though the author has not had access to the 2000 reprinted version in Lesotho orthography, all the editions up to 1974 (and possibly the 2000 reprinted version) fall short as regards coverage of current technological terms. Notwithstanding this, it is still the authority in Sesotho. As often

mentioned in literature, dictionaries such as Mabille's were targeted at second language speakers in the first instance and so, unsurprisingly, are bilingual in nature, mono-directional, and initially were mainly translation dictionaries suitable for use only by those much more functionally literate in Sesotho.

Their greatest disadvantage is the stem-approach for their macrostructure - entries - of major category words, which require a good knowledge of Sesotho morphology and, secondly, their article stretches which are confusing even for university students, for instance, a word such as *kgomo* 'cow' beginning with the complex sound [*kg-]* would only have its own article stretch and would not appear after *ke-* initial words such as *kena* 'come in' or k-initial words followed by the glide *[w-]*. In other words, the article stretches are not only alphabetic, but they are also arranged according to whether the initial sound is simple or complex, that is, phonemically, and so can be very confusing to the user.

The 1961 revised and enlarged 8th edition of Mabille and Dieterlen's dictionaries was reprinted in 1974 in the new spelling convention of the Republic of South Africa and titled *Southern Sotho-English Dictionary* in keeping with the name of Sesotho in the Republic of South Africa but bearing all the features of the Lesotho series. The culture of bilingual, mono-directional dictionaries - Sesotho to English or English to Sesotho – of varying sizes targeting different groups but mainly basic education level being produced by individuals continued into the 21st Century though trilingual dictionaries also surfaced. Except for Motsapi (2015), Pitso (1997) and Hlalele (2005), the dictionaries produced were smaller in size. Of these three, only Motsapi's is in RSA orthography whilst the other two are in the Lesotho spelling convention.

Dictionary name	Date of Publication	Edition No./ Revision	Author	Type and size
Southern Sotho-English Dictionary	1974	Only Edition	A. Mabille H. Dieterlen R.A. Paroz	Bilingual, 598 pp. in two column format and mono-directional
Bukantswe ya Maleme-pedi, Sesotho-Seafrikanse	1974	Only Edition	J.A. du Plessis J.G. Gildenhuys J.J. Moiloa	Bilingual, 269 pp.
Learner's English-Sesotho Dictionary	1985	Only Edition	Librarius	Bilingual, 1156 pp. one column format, mono-directional
Southern-Sotho-English Dictionary	1988	Only Edition	R.A. Paroz	Bilingual, 598 pp.
The Concise Trilingual Pocket Dictionary	1991	Only Edition	R.H. Moeketsi	Trilingual, 245 pp bidirectional
Sehlalosi: Sesotho Cultural Dictionary	1994	Only Edition	F.Z.A. Matšela	Bilingual 74 pp.
Khetsi ya Sesotho	1997	Only Edition	T.T.E. Pitso	Monolingual, 323pp.
New South Sotho Dictionary	1997	Only Edition	S.R. Chaphole	Bilingual 103 pp.
Sethantšo sa Sesotho	2005	Only Edition	B. Hlalele	Monolingual 325 pp.
Oxford First Bilingual Dictionary	2007	Inly Edition	D. Paizee	Bilingual 64 pp.
English-Southern Sotho Dictionary	2015	Only Edition	J. Motsapi	Bilingual 408 pp. two column Format, mono-directional
English Sesotho Official Foundation Phase CAPS Picture Dictionary	2018	Only Edition	Sesotho National Lexicography Unit	Bilingual 148 pp.

Table 2.7b: *Timeline and Survey table of Sesotho Dictionaries 1965-2018*

The first two monolingual dictionaries, Pitso (1997) and Hlalele (2005) appeared only after 121 years of Sesotho lexicography tradition began. Matšela's (1994) is a language for specific purposes dictionary (LSPD). On the other hand, the trilingual dictionary, Moeketsi (1991), is bidirectional from English-Southern Sotho-Afrikaans and Afrikaans-Southern Sotho-English and is strictly a translation dictionary with limited POS, syntactic and semantic information provided. Disappointingly, the two dictionaries developed in the republic of South Africa in 2015 and 2018 by Sesotho Government Bodies are also bilingual, having missed the opportunity to explore monolingual dictionaries. Such is the assailing power of English!

2.3.4. Dictionary making theories

In this research, a dictionary is referred to as a language reference resource and lexicography is approached from the school of information science. Information science presents new

concepts for user interaction (Heid, 2014). As an aspect of information science (Heid and Gouws, 2006), lexicography provides reference to linguistic data, it is however, not categorised as a sub-discipline of linguistics because it does not necessarily start at the point of studying language, but it contains linguistic information (Hartmann 1983: 13, McEnery 1992: 25, McArthur 1994, Heid 2011, Bergenholtz and Gouws 2012).

Learners presumably do not to hold complete comprehension of borrowed words when they continue to struggle with original words in their L1. The lack of comprehension means the incapability to detect factors such as ambiguity where a lexical item occupies more than one sense. However, the disadvantage of loanwords is that they do not allow for meaning analysis or morphological analysis. For example, *leksikhografi* 'lexicography' (Motsapi, 2016) cannot be decrypted using Sesotho morphological analysis where the prefix, stem, and suffix of the word are identified to deduce the exact meaning of a word. Therefore, the presence of loanwords in any language is a further call for a defining dictionary. Additionally, this indicates the need to understand the semantic theory of a language in lexicography and terminology creation.

Alberts (2001: 81) explains that "definition remains extremely important since it describes the meaning of words and terms and serves to standardise terms, especially in scientific and technical language." It is a valid point because definition illustrates the reason that has informed the choice to name a particular concept. Also, Alberts (2001) demonstrates that words are context-dependent and a monolingual dictionary as a type of a reference resource presents different senses of ambiguous words whose meaning determine their use in different contexts. On the aspect of assigning definitions to lexical items in the dictionary, Wiegand (1994) and Smit (2000) compiled "thematic question groups" as guidelines on aspects to consider when assigning meaning. These are as follows:

1. How can the (outer) form of the object be described? (Wiegand 1994: 272)
2. What are the "predicates characterising the hierarchically superordinate whole of which the object is an ingredient/part"? (Smit, 2000: 179)
3. In which way (how) does the object (type) come into existence? (Wiegand, 1994: 273)
4. What are the "predicates characterising activities in which the object functions/plays a role"? (Smit, 2000: 179)
5. Which other object is the object under consideration similar to and in which respects are they different? (Wiegand, 1994: 273)

Atkins and Rundell (2008) maintain that the dictionary as a source of data should reflect the full repertoire of language events for it to be complete. In as much as a dictionary compilation involves the user in mind, it should use labels that are comprehensible for the user. For example, when indicating that a phrase listed in the dictionary is a usage example, the abbreviation ‘e.g.’ as a commonly known abbreviation, can be used instead of ‘ex.’, which is incorrect although it can be argued that it stands for the first letters in ‘example’. Osselton (Hartmann, 1983:13) maintains that dictionaries first came into existence in response to practical needs. In contrast, Sesotho has a shortage of language learning resources, and school learners have educational needs that a First Language monolingual dictionary could fulfil. Rey (1995) adds that the social significance of dictionaries corresponds to the creation of an association between different languages. This research insists that particular languages should receive abundant terminological descriptions in and of themselves as a prerequisite to ensure accuracy in relating a language with others.

Podhajecka (2004) discusses issues encountered in dictionary-making and highlights dictionary typologies that influence metalexicography. The typology classifications are as follows: coverage, finance, access structure and target user (Podhajecka, 2004: 2). The discussion is mainly on the typologies that affect European metalexicography, but these are

important for this research for the classification of the dictionary. The classifications are according to the features of the dictionary and from this outline; the dictionary in this research falls under the informative dictionary classification.

Neubauer (1984) maintains that monolingual dictionaries are useful for second language learners, too, who have linguistic knowledge of the dictionary language. This indicates the usefulness of a monolingual dictionary as a language learning tool in that; once an individual knows the term/word in another language, they would aspire to know its meaning. This will additionally expand the vocabulary, competence, and performance in that language for L2 speakers/learners. Neubauer (1984) recommends that lexicography theory should be a guide in developing and improving practical lexicography work.

2.4 Dictionary types and classifications

There are various types of dictionaries that a language can have, and these include general, special, encyclopaedic, historical, linguistic and etymological dictionaries (Singh, 1982). They can further be subcategorised into monolingual, bilingual and trilingual dictionaries in terms of information presentation. Singh (1982) lists what is asserted to be the main criteria for dictionary classifications:

1) Density of entries: whether the word list is general or restricted and special? Does it also cover regional and social dialects, jargons and slangs and archaisms?

2) The number of languages involved: monolingual, bilingual, multilingual etc.

3) The nature of entries: whether lexical only or also encyclopaedic, the degree of concentration on strictly lexical data.

4) Axis of time: whether diachronic (dynamic) or synchronic (static).

5) Arrangement of entries: alphabetical or semantic or causal.

6) Purpose: whether normative or referential.

7) The prospective user: whether meant for the general reader to find out general linguistic information or for special users to know some special aspects of the lexical unit say etymology etc.? Is it meant for the general language or only for the language of literature, there too, the language of some author, here again the language of some of his works? (Singh, 1982)

Learner's dictionaries fall under Special dictionaries, whether monolingual or bilingual. Learner's dictionaries divided into native language learners and foreign language learners' dictionaries, and the latter is mostly translation/bilingual dictionaries. The density of entries for learner's dictionaries is at least 5000 words long. The nature of entries depends on the level or group which the dictionary is targeting. The purpose of learner's dictionaries is to aid language learner's in attaining information in a different language if the dictionary is bilingual, and to understand the properties of lexemes if the dictionary is monolingual.

There are three parts to dictionary-making, i) selection of headwords, ii) sourcing of data, and iii) the writing of the entries. Resources include other dictionaries, citation files, computer corpora. "Computer corpora can be searched rapidly and efficiently. It can be used for checking information, or for seeking answers to specific queries. However, more significantly, it can provide raw data for the construction of dictionary entries" (Jackson, 2002: 29). Dictionaries vary in size, dimension, length, and coverage. The Sesotho dictionary that is compiled is aimed at being concise. Although the dictionary-making tool, *TLex* allows for the recording of etymological details this aspect is not thoroughly explored herein.

A dictionary has a macrostructure and a microstructure. The macrostructure is the front matter of a dictionary, the body and the appendices. The front matter includes an introduction or preface, (explaining the innovations and characteristics of the edition concerned, a guide to using the dictionary), which may consist of a single page diagram. One of the aspects that have been observed from Sesotho dictionaries, the bilingual ones, is that they only explain the front

matter in English or Afrikaans, which further suggests that these dictionaries were not created for L1 Sesotho users. The latest example of this is seen in The *Official English-Sesotho Dictionary* (Motsapi, 2015) that was compiled and published under the Sesotho National Lexicography Unit of PanSALB. Other front matter might be an explanation of the transcription system used to indicate pronunciation, a list of abbreviations used in the dictionary and an essay on some relevant topic, such as the history of the language or varieties of the language around the world.

On the other hand, the microstructure of a dictionary refers to the arrangement of the information within the entries. In terms of microstructure, monolingual dictionaries are written in one language about that language, and they can contain definitions, parts of speech, etymology, encyclopaedic, pronunciations (audio for electronic dictionaries, or transcribed pronunciation), usage examples, and pictures depending on the purpose of the dictionary. Bilingual and trilingual dictionaries on other hand are primarily translations of words between different languages; they may or not have parts of speech annotations, may or may not include pictures.

In terms of dictionary traditions, all modern-day dictionaries are written using computers and therefore, start-off as electronic dictionaries even though some may further be converted into print while others are transferred to online platforms. An online or electronic dictionary is, "A dictionary whose data exists in electronic form and can be accessed through various different electronic media" (MacMillan Dictionary, 2019). An online dictionary can be accessed from the World Wide Web or can be installed on electronic devices in the form of computer applications. Such dictionaries are primed with human languages using computer/programming language scripts. There are dictionaries online that are easily printable, and there are those that cannot be printed in the form of a traditional dictionary book. There are different computer scripts or software used to create exclusively electronic dictionaries and

electronic dictionaries that are printable. Volume II of the *Bukantswe* (Oliver, 2006), whose first volume is not traceable, is itself no longer traceable online. It was in a fixed tabular format. A user did not have to key in a word to generate output because the words were already listed on the interface just as is the case in a printed dictionary. One would scroll down a page to find a word. Its only flexible feature was that it allowed the user to search starting either from English to Sesotho or from Sesotho to English.

With Volume III, on the other hand (Oliver, 2009), it was designed on a Personal Home Page (PHP) script. PHP is a processor that uses server-side scripting and as a result, does not produce an output of the direct command. The dictionary provides an output based on word stems; it presents *all* results that have within them the queried word or its stem. These results, which are out of context, sometimes are phrases instead of words. It does not prioritise the searcher's command in that the actual query will be found below other entries that contain the same word. For example, in Figure 2.4. below, the word searched was *motho* 'person', yet the outputs either describe types of persons or their actions, none of which gives the equivalent of what 'person' is. It is observed that structure determines whether a dictionary is printable or can only be accessed electronically.

Figure 2.4: *Online Sesotho - English dictionary – Bukantswe ya Sesotho le Senyesemane*

The *Hiswelókë's Sindarin Dictionary* (HSD) Project (2007) lists three approaches to online dictionary-making, and these are a typographic, editorial and lexical approach. Example of a typographic approach is a two-dimensional printed page, and the editorial approach is one-dimensional. The lexical approach, recommended, allows for continuous encoding of the dictionary. Having a typographic and an editorial dictionary format is as good as a printed dictionary because adding new entries to it will be equal to manually starting a new dictionary. The HDS Project (2007: 1) highlighted that,

> The main interest in having well-formed electronic dictionaries, in our opinion, is that other compilers and scholars should be able to almost automatically extract pieces of information from the initial document and present them differently - and even possibly use the same information to produce a wholly different work, not envisioned as a possibility by the initial compiler.

This is achievable when looking at the lexical aspect of the dictionary, i.e. the way the underlying information structure is to be encoded. The lexical view includes the underlying information represented in a dictionary, without concern for its exact textual form. The advantages of some electronic dictionaries are that they can be modified, and space is not a concern as it is with paper dictionaries depending on the computer language script used to write the dictionary.

2.5 Chapter summary

This Chapter has reviewed the literature with respect to the historical background of the language, its morphology and lexicographical tradition. The latter has indicated that on the main, bilingual dictionaries have been the preferred choice and that existing online dictionaries are still rudimentary in Sesotho lexicography. It is the aim of this study to address these deficiencies to meet the current social needs of Sesotho mother language speakers. As lexicography deals with words, a review of the morphology of Sesotho was undertaken to

inform the decisions on the compilation of the envisaged learner's dictionary. The next chapter will outline the methodologies and methods that will be applied in Chapter 4.

CHAPTER 3: METHODOLOGIES AND DATA COLLECTION

Introduction

The aim of this research is to design a corpus-based monolingual Sesotho learner's dictionary, and this chapter presents the methodologies adopted and the methods that were used to collect data by demonstrating how the corpus was compiled and processed to generate the lexical items for the monolingual dictionary that is the subject of this research. The chapter also discusses the design of the dictionary by giving a systematic reflection on the concordance tools, the dictionary development software, and additional computer programs that were used to process the data. The following section is a detail of the research design that determined all the activities undertaken towards achieving the aims of this research as well as addressing the research questions.

3.1 Methodologies

The first methodology adopted in this study for the compilation and analysis of data, is the Corpus Linguistics methodology. Corpus Linguistics, henceforth, CL, is an approach or model of linguistics that studies the structure of language from its different components such as phonetics, phonology, morphology, syntax and semantics and language use, that is, sociolinguistics. Further, it is also used in allied disciplines such as Lexicography, Terminology, Translation Studies etc. It is not an aspect of linguistics but is applicable to all its aspects and allied disciplines, hence corpus-based phonetics, corpus-based morphology, corpus-based lexicography etc. as opposed to non-corpus-based phonology or non-corpus-based semantics (McEnery and Wilson 2001:2). McEnery and Wilson (2000: 1-20) trace the history of CL and conveniently organize it into two eras; Early CL and Contemporary CL. The period before Chomsky's revolutionary criticisms (mid-1950's) on the methods used to collect data, the nature of collected data and the skewed results arising therefrom is referred to as Early

CL, which was better known as *field linguistics* then and which Chomsky considered basic or primitive.

Field linguists and later structuralists compiled large collections of utterances from the language under investigation and analysed it carefully in steps using the bottom-up strategy. Their data was not called 'corpus' then but it underpinned their methodology and in contemporary times it is considered 'corpus like' (McEnery and Wilson 2001: 3). In the wake of Chomsky's criticism, this framework was abandoned - though not completely - and the new framework, which is called CL. The period after Chomsky to date - came into being and ushered the contemporary methodology for linguistic analysis of data, now known as 'corpus linguistics' whose unit of analysis is a corpus, which will be explained in more detail in 3.2. Since post-Chomsky, corpus is now both written and spoken and currently includes multimedia in all its forms and uses. Corpus collection in contemporary CL is empirical in nature. Even though CL received criticism in the 1950s (McEnery and Wilson: 2001), it has since proven highly effective, efficient, and reliable wherever it is used. Kennedy (1998) asserts that CL is a significant tradition in language practice and provides a convention for the compilation of text samples into a corpus that can be analysed or used when investigating language.

CL uses either or both quantitative and qualitative approaches to analyse its data, that is, corpus depending on the nature of the investigation. This research relies on both approaches. The quantitative approach is statistics based; it focuses on numerical data. The results processed using this approach have to be measurable or rateable. In this research, this approach was applied through sampling and conducting word frequency analyses, which assisted in deciding which words to select for use in the dictionary. By processing the corpus data to get frequency counts, the arranged results were then easily comparable, which made data deduction and transfer from the corpus tool to the dictionary tool uncomplicated.

The qualitative approach on the other hand, entails doing research by collecting and analysing data such as text in order to understand concepts. It is used when generating new information from observation of already existing data for further insight into the concept being studied. This approach applies to the study of language because it is useful for textual analysis. In this research, this meant studying the structure and function of lexical items. The advantages of the qualitative approach in this study include the opportunity to study natural language in different contexts; observing how a single word was used in different texts and how its different uses affect its meaning(s). One of the disadvantages of this approach is that the results are not conclusive, this is particularly true in the case of studying words and their meanings because the data collected may not entirely represent all areas in which the language is used, and words are often understood differently by different speakers. "Qualitative research is characterised by its aims, which relate to understanding some aspect of social life, and its methods which (in general) generate words, rather than numbers, as data for analysis" (Patton and Cochran, 2002: 2). Patton and Cochran (2002: 3) add that qualitative research methods receive criticism on the basis that, "samples are small and not necessarily representative of the broader population, so it is difficult to know how far we can generalise the results."

There are two major approaches to CL, one is the corpus-based approach (CBA), and another is the corpus-driven approach (CDA), and in both, the unit of analysis is a corpus. CBA is an empirical approach wherein data patterns are observed. It does not rely on rational information inquiry, and it is used to test and improve language study. "Corpus-based approach research assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of the research is to analyse the systematic patterns of variation and use for those pre-defined linguistic features" (Biber, 2010: 206). Biber (2010) adds that CBA studies language from naturally occurring texts that is, both written and spoken.

Tognini-Bonelli (2001:84) defines the second CL approach, corpus-driven approach, CDA, as a method that studies a language corpus without paying attention to linguistic rules but makes conclusions or claims from corpus observations. In essence, CDA concentrates on the "observation of certain patterns that lead to the research hypothesis" (Tognini-Bonelli, 2001: 17). Also, Tognini-Bonelli (2001: 84) asserts that a CDA is a method, which "builds up the theory step by step in the presence of the evidence". In other words, CDA starts with research before formulating a theory, but it focuses on the corpus data to determine descriptive linguistic information. It also aims to build an ideal plan without any considerations of the corpus theoretical principles (Gast, 2006: 114). Further, "corpus-driven research is more inductive so that the linguistic constructs themselves emerge from analysis of a corpus" (Biber, 2010: 1).

Additionally, implementing a corpus-driven analysis is not reliant on the size of the corpus, what is essential are the results revealed from the corpus (Leech, 1991; Krishnamurthy, 2001; McCarthy & Carter, 2001). On another hand, Pienaar & Klerk (2009) assert that a corpus size is always determined by socio-linguistic factors relating to the population under consideration and by two, the purpose that the corpus will serve.

Both CBA and CDA are used to analyse language structure and use (Biber, 2010), both use a corpus as a unit of analysis, the nature or characteristics of corpus and methods of corpus compilation are the same. However, they differ in their assumptions as regards theory. As CBA is the oldest of the two approaches, it has been used extensively to create corpora for different languages. Using corpora for lexicography purposes is a method that was pioneered by John Sinclair, and one of the first uses of corpus in lexicography was the development of the *Collins COBUILD English Dictionary* of 1987, which "was the first dictionary ever to be based on a corpus of real language data, so words that were included were current and attested. It simplified the layout and typography of dictionaries by not using any abbreviations or symbols

in the main text. It also gave copious examples of authentic language use taken directly from the corpus" (Krishnamurthy, 1987). The *Oxford English Dictionary* project (Kennedy, 2014: 11) also used a corpus for dictionary making, and this project continues to expand both the English corpus and create different lexicography products. For African Languages lexicography, corpora have been used by the likes of De Schryver (2007) for the *Oxford Bilingual School Dictionary: Northern Sotho and English*, for example.

As far as compilation of the corpus is concerned, both approaches, which are inherently similar except for their assumptions and point of departure, were used. However, CBA was used for the development of articles of the dictionary. Herein the corpus is a tool from which the lexical variations from different contexts were studied, and then used to assign usage examples to lexemes in the dictionary. The methods relating to these approaches will be discussed in Section 3.2. below.

In addition to using Corpus Linguistics in this lexicographic study, the related approaches to dictionary development that were studied are the Semasiological and Onomasiological approaches to lexicography. Semasiology is a dictionary structural frame that presents dictionary data from the direction of concept to meaning, while Onomasiology studies meaning to assign to concepts. With the Semasiology approach, the linguistic properties of a word are essential for a dictionary. These properties include the meaning of the word (semantic property), usage of the word (syntactic), the pronunciation of the word (phonologic), the word variation (morphologic), and the aspect of contextual examples, which is a factor of pragmatics. This research adapted the Semasiology approach together with CBA and CDA in the design of the dictionary.

3.2 Corpus

In Corpus Linguistics, data is corpus or corpora in plural, which derives from the Latin equivalent for body, which is the basic unit of study for the application of CBA and CDA methodologies. Therefore, in contemporary CL, a corpus is a body of machine-readable text samples -written or spoken - of acceptable or standard size; the texts of which are a balanced, representation of a language being investigated. A corpus can be fixed to a certain number of words, closed, or be open-ended i.e., can be added on continually until the target size is reached. A corpus has two parts: a header and a body. The header is annotated with extralinguistic or encyclopaedic information which is essential for the identity of any given text while the body is tagged with linguistic information. Parts of Speech (POS) or morphological tagging is the most popular linguistic annotation and is essential, especially, in lexicography.

3.2.1. Nature and size of compiled corpus

The corpus collected for this study is a monitor corpus, i.e., it is open-ended. In early post-Chomskyan corpus studies, a corpus was regarded as a representative standard of a language if it amounted to one million words or above (Herbst, 1996; McEnery and Wilson 2001) such as Brown corpus, Lancaster-Oslo-Bergen, International Corpus of English etc. However, there have been corpora of less than a million words such as the Helsinki Corpus of English (245 000 words), Helsinki Corpus of Early American English (500,000 words), Augustan Prose Sample (80,000) etc., which have been useful for research. However, the concept 'size' has changed over time due to the advances in computers. The British National Corpus, popularly known as BNC consists of a 100 million words while the Chinese zhTenTen monitor corpus whose initial target was 10 billion words and based strictly on internet resources is made of 15.9 billion words as of 2017 https://sketchengine.eu/-chinese-corpus/.

Though the concept 'standard' has changed with the advancement of computers, it has not changed in terms of genre coverage. From a survey of the size of earliest corpora, it is not amiss to conclude that even a small corpus can have very good use. The corpus compiled for this study is 101 359 words and so does not meet what was previously deemed the standard size of a corpus in the wake of contemporary Corpus Linguistics in the 1960s – 1970s, which was one million large corpus as already mentioned earlier. The corpus used herein is smaller because it is a provisional corpus. De Schryver and Prinsloo (2000: 98 – 99) maintain that "A corpus of just 300,000 words does not even come close to the going rate of one million words. Yet, there are at least two good reasons to support the compilation of small-size corpora as a first approach. Firstly, from a theoretical perspective Kennedy note:

> The new generation of mega-corpora is likely to dominate work in corpus linguistics for a considerable period. However, while definitive lexical descriptions and large commercial projects will be crucially dependent on them, such mega-corpora may not be appropriate or easily accessible for individual academic researchers [...] It is important therefore not to overlook the value of working with small corpora of one million words or less in the period ahead [...] There is nothing magic about one million words and no advantage in compiling a corpus of that size if a corpus of 200,000 words will do the job intended (Kennedy, 1998:56-57, 73).

Considering Kennedy (1998) and De Schryver and Prinsloo's (2000) assertations, the 101 359 words corpus is appropriate for this study, which verifies these assertions. Although, Sampson and McCarthy (2005: 83) warn that a small corpus "… will be adequate for the study of the fairly frequent patterns and meanings of many thousands of words but will not be adequate for a reliable description of the language as a whole." Nonetheless, without dismissing the caution, this would apply if the corpus was compiled for the purpose of a grammar book because in that case it would be essential to describe the language in its entirety.

3.2.2. Machine readability

Although machine readability was not always an important feature of corpora in early corpus studies, it is now a key standard for a corpus that is used for research or application purposes. A corpus is machine readable if it can be processed using computational language processing tools. Therefore, the physical or hardcopy texts that were collected for this study were scanned and converted into the appropriate file format, and those that were unscannable typed into computer-readable format. One of the biggest advantages of electronic documents is that they require very little storage, are not easily erasable, are portable, exchangeable, shareable, processable on concordances and are available at a click, virtually, anywhere.

3.2.3. Sampling, balance and representativeness

There are six text types represented in this corpus; from random samples of fourteen text genres and their writing periods vary from the 1830s when Sesotho was first reduced to writing by the missionaries to 2015. The corpus is diachronic in this manner in that it reflects how Sesotho has changed over the years, for example, the 1830's texts were all written in what is now considered Lesotho orthography because South African Sesotho orthography changed in 1961 by dealing away with diacritics and reformed semi-consonants and some consonants (see Section 2.1.2. The Orthography of Sesotho). Only written texts are included in this corpus. It is noteworthy that even though there is a news bulletin included, it was not recorded and transcribed by the researcher. Instead, a newsreader provided the text in its primary written form. Using spoken texts would have required transcription, which would have required additional time than available for the size of this research, and software of its own to ensure accuracy.

Compiling the corpus for this study included the collection, selection, sampling, sorting and processing of texts. The c.a. 100 000 words were processed from a selection of different

Sesotho texts. Most available Sesotho texts are from the literary genre; as a result, novels, short stories, poetry, publications of transcribed music, and dramas were collected. Additionally, the Sesotho version of the bible and Sesotho hymnals which are the early Sesotho written text were sampled. Articles from a Sesotho 1930s newspaper, 21st century News bulletins, and parliamentary speeches, advertisements, medical surveys, a Sesotho High school Activity book and past exam papers were also gathered. A detailed analysis of these texts is provided in Chapter 4.

Sample amounts depended on the number of texts available for each text type. For example, only one Sesotho newspaper was available at the time of the collection, and as a result, 60% of articles from this newspaper were sampled. On the other hand, of the four different novels collected, 20% of each book was sampled. Selected texts are in both the Republic of South Africa (RSA) and Lesotho orthography because there are genres that were not available in RSA Sesotho spelling such as parliamentary speeches and newspaper articles; there was also no RSA Sesotho newspaper in South Africa at the time of corpus collection. The newspaper that has been sampled in the corpus is written in Lesotho orthography and it was accessed from a South African library archive. Most of these texts were only available in hardcopy because only a few are online. The hardcopies such as novels, were accessed from the University of the Witwatersrand Libraries, and the Activity book and past exam papers were shared with the researcher by a Sesotho teacher from Tshepang High School in the Clocolan Township in East Free State, where 86% of the population speaks Sesotho (Census, 2011).

In corpus building, balance in the types/genres of texts included in the corpus is considered one of the features that make for a well-designed corpus. Balance is the state of having a representation of various text genres in the corpus. The rationale being that when

various genres are included, it allows for a study into the different structures and uses that single words can occupy further allowing for broader interpretations or analyses of words, and this is critical information for a dictionary. Xiao (2007: 12) maintains the following in terms of balance in a corpus:

> A balanced corpus covers a wide range of text categories which are supposed to be representative of the language (variety) under consideration. The proportions of different kinds of text it contains should correspond with informed and intuitive judgements. There is no scientific measure for balance – just best estimation. The acceptable balance is determined by the intended use – your research questions.

Biber (1993) defines representativeness as "the extent to which a sample includes the full range of variability in a population. In corpus, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness." Representativeness was achieved in the compilation of the Sesotho corpus by collecting texts from the 1830s to 2015, and these are in various literary genres. Because there was no target population in compiling the corpus, it is a general corpus intended to represent the language, although because of its size, cannot be said to be maximally representative.

3.2.3.1. General literature corpus

Most Sesotho texts are from literature. Sesotho novels, drama, poetry and music are easy to access more than any other type of writing. Additionally, the literature genre is worthwhile because it is always in production, meaning that these texts are a credible resource to use to trace how the spelling and orthography of Sesotho have changed over the years. Therefore, there are more literature texts in the corpus than any other genre. The genre is further convenient because it covers different subjects, and its lexicon is varied. There are novels from the 1980s as *Mopheme* (1982) and *Seyalemothati* (1981) for example, which are very popular texts in Sesotho literature to this age.

3.2.3.2. Religious and political texts

The bible and hymnals are the earliest Sesotho written texts in Lesotho spelling, and these are still used in modern society in their original form. The religious genres make a reference to a world of its own; it tells stories that perhaps a learner born in the 2000s would not begin to perceive and thus defining texts such as these clarifies the description of a world that one may not have been exposed to or can relate to due to the use of metaphoric language in some instances.

The political genre was also considered because of it uses a sub-language, its own jargon that is somewhat distinct from everyday language. The genre shows how language is used in a formal context. However, the challenge with texts from this genre is that they often contain English words. The assumption is that this is done by the authors of the texts because there are still political terms that have not been assigned Sesotho equivalents and perhaps even if an author were to translate political terms and concepts, the challenge is whether their audience/laymen would understand the translated terms as intended.

3.2.3.3. Educational, Medical and Commercial corpora

The education genre should be the genre with most samples, but the available texts in this genre are mostly literature texts, thin activity books and examination papers. The word types used in activity books and examination papers, which are written in South African orthography, seem the same throughout higher grades of school. The structure of questions, the elaboration of situations, and the instructions usually take the same form in textbooks used between Grade 10 – 12. In other words, the writing style is the same; hence the decision to use only a select few education/curriculum texts.

The medical and commercial texts were provided by a freelance proof-reader from a South African-based Translation and Localisation company, *Lingua World*. The terminology in this genre is relevant for a learner's dictionary because it relates to school subjects such as Business Studies and Science, and the sampled texts contain technical words as can be expected. All the texts have been published and were made available for public consumption. Therefore, they are credible sources that can be said to have undergone some sort of verification by language experts.

3.2.4. Annotation

Annotation is one of the characteristics of a standard corpus; it refers to the assignment of information regarding texts or contents of a text. It applies at word or sentence level and the extralinguistic information of a document. As indicated in 3.2. above, a corpus has two parts, which are the *header* and the *body* both of which, when annotated, can make a text easy to identify and make information retrieval and analysis smooth as well. There are conventional standards of corpus annotation. One of the earliest is the *Helsinki* convention with a 25 tags annotation scheme only for the header of a corpus. There is also the *Text Encoding Initiative* (TEI) scheme which is a vast improvement of the *Helsinki* and other schemes, meant as an international standard and capable of annotating both the header and the body. However, because it is too complicated, a simpler version of it was developed and labelled *TEI-Lite* (McEnery and Wilson: 2001), which can provide both linguistic and extratextual information for each single text sample. Though on the one hand *TEI-Lite* has the capability to assign labels or tag word tokens in a text morphologically on the other hand, body annotation is language-specific and may require the tag set within TEI to be modified to suit the language in question. Tag sets are useful in automatic corpus annotation because they allow, for example, a computer programme to identify words and their appropriate lexical usage. The corpus used herein is

raw, that is, unannotated due to the paucity of tag sets suitable for South Eastern African Languages.

Notwithstanding, headers were developed for all the texts in this corpus using *TEI-Lite* in order to facilitate their sorting into different genres among other factors. A description of each text was done using the *TEI Lite* descriptions. TEI guidelines were considered because TEI is regarded as a flagship in corpus annotation (McEnery and Wilson, 2001:35), and because there is currently no standard way of representing information. Additionally, TEI guidelines are credible as they have been adopted by large corpora projects such as the BNC (Bumard: 2000).

3.3 Tools for data processing and dictionary design

These tools are classified into two categories: tools for compilation of corpus and tools for dictionary making.

3.3.1. Corpus compilation Tools

The tools used to compile the corpus are the following: a concodancer, photocopier and scanner and they are discussed in this section. As stated in 3.2., the data used herein had to undergo certain processes in order to use it for the creation of the dictionary. With the corpus being a tool itself, it was processed using technologies such as scanners, text conversion tools, concordancers, etc. A number of tools were experimented with before the preferred choices were decided upon and this was for both data processing and dictionary design.

In order to create a wordlist from the corpus, a couple of tools were explored such as *AntConc*, *Wordsmith tools*, *LexiquePro* and eventually, *TLex.* The first tried tool was *AntConc* (http://www.laurenceanthony.net/software/antconc/), a corpus analysis tool that is freely available online and was first developed in 2014 (Anthony, 2014). Its built-in features include

a concordance, concordance plot, file view, clusters, collocates, word list and keyword list. It can handle large amounts of words and would have been handy in this research, however, it does not have an integrated dictionary making interface like the software that was eventually settled with for both corpus processing and dictionary creation.

The second explored text analysis tool, *Wordsmith Tools Version 6.0 (*http://lexically.net/), is a language processing application - concordance - that has three mechanisms; (i) *Concord*, a concordance tool that retrieves strings of characters from the collected corpus, (ii) *Keywords* generator that is used to identify keywords in different contexts; and lastly, (iii) *Wordlist* program feature, which is a software designed to generate a wordlist in an alphabetical order, and frequency based list of words from text data (Scott, 2010). Initially, *Wordsmith* was to be used in this study to generate a wordlist and further study keywords, i.e. lexical items selected for the dictionary in their contexts in order to retrieve usage examples. However, *Wordsmith Trial Suite* does not have the capacity to process more than 50,000 words at one go, and the corpus collected is over 100 000 words. The full suite is patented and initially there were no funds to purchase it. As such, one would have to upload multiple portions of the corpus one after the other, which would not ultimately give wholesome results. This was infeasible in terms of time.

*Lexique Pro 2004 (*http://www.lexiquepro.com/index.htm) is a dictionary software that is freely available, and it does not have language or lexicon distribution restrictions. It is suitable for both online and printable dictionary creation. Some of its outstanding features include audio and image display. With *Lexique Pro*, one can create monolingual to multilingual dictionaries that contain attributes such as synonyms, variants, morphemes, main entries, cross-references as well as sub-entries for individual entries. Bowern and Dwyer (2012) regard *Lexique Pro* as the top suitable software for dictionary making; however, the limitation of this

software is that it does not include any text analysis features such as concordances or wordlist generator.

Fortunately, an all-encompassing dictionary making tool was discovered, and this is the TshwaneLex *(TLex)* Suite, a dictionary compilation software, which includes a built-in concordance, the *tlCorpus* (https://tshwanedje.com/corpus/). The economical decision was to use it for both text processing/analysis and dictionary making and, fortunately, by then some money had become available to the Wits University department of African Languages and Linguistics and the researcher managed to purchase a personal license after losing access to the university license. Although *TLex* requires a licence, it offers all the corpus processing and dictionary-making software in one suite and was a better buy. *TLex* includes comes with a concordance feature, *tlCorpus*, which is different from *Wordsmith* and *AntConc*. However, like the other two, *tlCorpus* allows for the generation of word count frequencies, processing of keywords and wordlist generation.

3.3.2. Dictionary making software: *TLex Suite*

As mentioned above, *TLex Suite* is a multifunctional dictionary-making suite, which allows for the creation of monolingual and bilingual dictionaries, and it is adaptable to South African indigenous languages. It was developed by the TshwaneDJe Human Language Technology software production company, who introduced it in 2003 (Joffe et al.,2003). This suite currently includes a dictionary compilation software, *TLex*; a terminology compilation software, *tlTerm*, a concordance software; *tlCorpus*, and a database analysis application, *tlReader* (https://tshwanedje.com/tshwanelex/).

TLex is query-based and in this way minimises the time spent on word searches by the end-user. By contrast, most online dictionary assessments look at the speed that a tool takes to

process results; *TLex*, for instance, has a quick search query function that gives a direct response to a user's exact query. Additionally, De Schryver and De Pauw (2007) point out that *TLex* is exceptional because of its integration of a Dictionary Writing System (DWS) with a Corpus Query Package (CQP), which makes it agreeable for a corpus-based approach to dictionary-making. This feature recognises that a web (electronic) dictionary should have the ability to do things that cannot be done on a print dictionary (Bowern and Dwyer, 2012). By integrating the DWS with a CQP, *TLex* allows one to import linguistic information into the dictionary interface.

It presents dictionary entries with their POS tags alongside definitions, and usage examples, among other features. It is one of the prominent dictionary-making tools now for South African languages. It is customizable, reliable, can feature images and sound, multiple exports, and most importantly for this account, it has built-in facilities to import corpus data such as its *tlCorpus* application. The default language of *TLex* is English, which means that the annotation codes are in English, but it is possible to configure them to other languages. Further, *TLex* is a commendable software as it additionally allows for onomasiological and semasiological data presentation.

Unlike the method used in the already existing Sesotho online translation dictionaries, which is keyword based, *TLex* gives only the results of the query and no other words with the same root/stem as the word searched for. Moreover, with *TLex*, once the dictionary is online, it allows for smooth user interactions and the designer can always restrict access when making configurations.

3.3.3. Photocopier, scanner and camera

The hardcopy texts were photocopied and scanned using various light production printing machines from the Wits University libraries. Only sampled sections per text were copied and scanned. All printers had photocopy and scan features, and because they were colour-laser printers, the visibility and colour of texts was not compromised. In the case of the only Sesotho newspaper sampled for the corpus, the 1930s newspaper, the articles were compiled by taking camera shots as it was advised by the library that the newspaper was too fragile and therefore, could not be ran through a scanner, and a phone camera was used for this.

3.4 Methods/Steps

In this sub-section methods used to build the corpus and to develop the dictionary are outlined.

3.4.1. Corpus compilation

This process involved the following steps:

- Identifying appropriate texts/determining scope and size of corpus
- Identification of texts sources
- Collections of texts from physical libraries, the Department of Education Website, Lesotho Parliament Website, and online literature blogs.
- Sampling of selected texts
- Scanning of hardcopies
- Copy-pasted softcopies from the web
- Took camera shots of the newspaper articles
- Post-editing; conversions of scanned copies into *Word* documents
- Re- typing of data from other hardcopy texts
- Creation of Header Annotations for each text identification

- Combination of all texts into one *Word* document, adding up to a corpus of 100 000 c.a. words

3.4.2. Steps for dictionary design

- Naming the dictionary
- Selection of the dictionary type: monolingual
- Customised *TLex* for Sesotho by changing the default language for POS and cross-reference markers
- Upload of the corpus onto the concordance – *tlCorpus*
- Generation of a wordlist using *tlCorpus*
- Download of the *tlCorpus* wordlist, which also indicates frequency counts per item on the wordlist
- Identification and deletion of foreign words from the Wordlist
- Sifting out numbers and other foreign characters
- Selection of the high frequency words
- Transfer of high frequency words onto a *Microsoft ExcelSheet*, which assisted with counting the number of words that were processed on the Wordlist
- Transferring the list of 5000 words from *Excel* to *TLex* interface

3.5 Scope of study

Though 5000 words were identified for the target dictionary, only 250 (Annexure 1) were selected for inclusion in this study in consideration of its restrictions. Consequently, headwords and articles under A and B form the basis of this investigation. More work still needs to be done in terms of assigning meaning to the rest of the +4 750 words, and this will continue to be done post this research. After completing the steps in 3.4., articles were created for each

lexical item of the 250. Each article contains the word, meaning, POS, usage example, etymology where applicable, and further, pronunciation.

3.6 Summary of chapter

This chapter outlined the methodologies, methods and tools used for the primary aim of this research. The methodologies were discussed, the tools described and the methods were listed. CL methodology assisted in crafting the approaches to consider and apply in this study. The corpus, as a unit of study in CL and a tool for this research, contributed to the contents of the monolingual Sesotho dictionary, and this shall further be demonstrated in Chapter 4. Chapter 4 illustrates how the methods and tools discussed here were applied to analyse the data in response to the aims of this research.

CHAPTER 4: DATA ANALYSIS AND RESULTS

Introduction

The aim of this chapter is to present the analysis of data collected using the methods and tools arising from the selected methodologies outlined in Chapter 3 in order to respond to the research aim and questions of this study, which is to design a Sesotho monolingual Learner's dictionary using a corpus-based approach. The first step towards achieving this aim was to process the corpus using the *tlCorpus* concordance tool to generate a wordlist i.e., the data, and then analyse the distribution of words per POS as these are the bedrock of the macrostructure of the dictionary being developed. The example sentences were selected from the corpus for its microstructure. The second phase was to customize and configure the *TLex* interface and its settings to suit the design of this kind of dictionary. Further, the analysis demonstrates how the lemmas were marked-up. In the process of analysing the data, results which are inevitable, are also presented. The chapter is divided into four main sections: corpus analysis, customization of the interface of the dictionary software, dictionary design and results discussed in that order.

4.1 Corpus Analysis

The corpus of 101 359 words was harvested from different written text genres as shown in Table 4.1. below which illustrates the text types or genres included in this corpus. In sampling the below texts, the aim was to prioritise texts that learners currently use in school, which includes novels, poems, activity books, and exam papers. By doing so, the dictionary contains words that are aligned to the learner's comprehension level. As mentioned in 3.2.3. Sampling, balance and representativeness, these texts came from physical libraries and online archives. The hard copies texts were sourced from libraries, and they include the novels, poetry, traditional

music and newspaper articles. Some of these were photocopied, others scanned, a few taken photos of, and some portion was manually typed into the computer from the hardcopies.

Text Type	Genres	Year of Publication	Author	Region
General Literature	Novel Chapter Short Stories Poetry Traditional Music	1982 2014 2011, 2012	S. Matlosa M. Mahloane D. Kunene, Unknown	Lesotho South Africa Not Specified, SA - JHB
Religion	Hymnals Bible Chapter	1830s	Compiled	Johannesburg
Socio-Politics	Parliamentary Speeches News Bulletins Newspaper Articles	2007-2012 2013 1939	Various Authors L.Murobane Motheo FM Mochochonono Newspaper;	Lesotho Lesotho, Maseru; SA, Botshabelo
Education	Activity Book Novel Poetry Exam Question Paper	2011 1987 1992 1993 2014	M.E. Mofokeng, E.M. Skosana, M.D. Ntjhoboko, S. Leche R.J. Masiea K.P.D. Maphalla K.P.D. Maphalla	South Africa
Health	Medical Surveys	2015	Various AUTHORS	Pretoria

Table 4.1: *Texts sampled for the corpus*

The corpus used is a monitor corpus, that is, open-ended and allows for the addition of more texts on a continuous basis (McEnery & Wilson, 2001) and consist of written sources and transcribed speeches. The corpus is not annotated; it is diachronic as the samples were extracted from various texts written during different periods i.e., from the 1830s to 2015. Further, the

texts are written in two different Sesotho spelling conventions: RSA and Lesotho, and they include borrowed words, dialectical varieties of words and there are instances of code-mixing.

The Parliamentary speeches and novel samples have the highest coverage in the corpus in terms of number of words extracted from them and combined they make up 46% of the corpus, followed by medical surveys at 14%. As indicated in Figure 4.1. below, the activity book and traditional music have the smallest percentages with the exact word count for the traditional music, for example, being 2027 words as the book itself is small in size unlike novels.

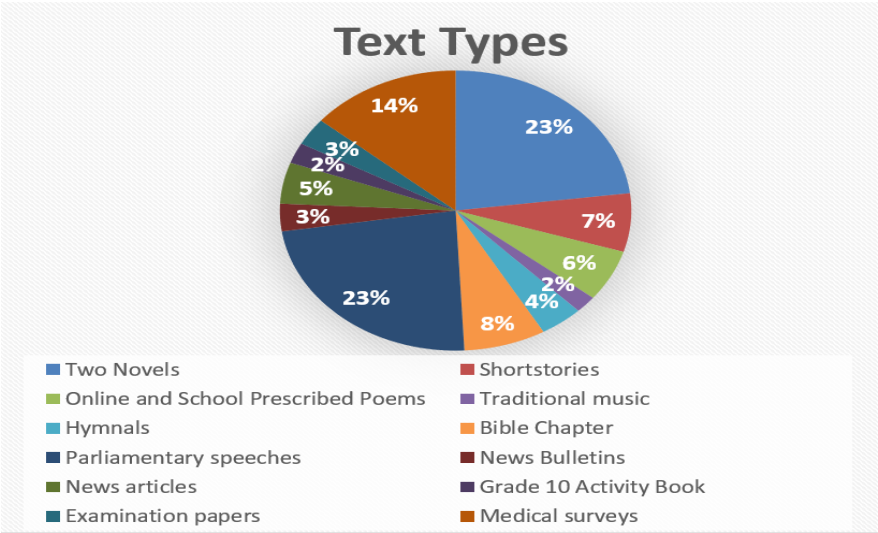


Figure 4.1: *Distribution of text genres in the corpus*

Novels have 40 000 – 100 000 words as a standard requirement. Even though it was desired to have a large amount of school texts in the corpus, only one activity book was added to the compilation and most other school texts were novels and poetry anthologies. The decision to sample from one activity book was made after discovering that spelling and grammar mistakes abound in textbooks and, therefore, to eliminate excessive post-editing only one was selected.

The speeches were recorded and saved in *PDF* format on the Lesotho government website (https://www.gov.ls/document-category/official-speeches/) and could be copied and pasted onto a *Word* document without converting the *PDFs* into *Word Doc.*, and no information was lost in the process of copy-pasting. The total word count for the speeches is 23 978. The novel chapters were scanned and converted into *.doc* format as required by *tlCorpus* for processing, and some were typed manually. The total word count for each of the two novels is 20 437 for *Mopheme* and 2 511 *Seyalemothati*. The medical surveys were sent to the researcher via e-mail by a Sesotho-English translator who had translated the surveys as a freelancer a medical research organisation.

The short stories and some poems were retrieved from online sources. It was observed that poems retrieved from websites and those taken from the school prescribed texts carried the similar themes of protest, praise and imagery. There is no guarantee regarding whether the poems that were accessed online had been screened to ensure appropriate use of language because they were taken from web pages created by independent writers. However, the writing was of good quality and well-edited.

The news bulletins samples were sent to the researcher by a community radio journalist and the text presented some grammar errors and sub-standardized spelling. As a result, the text was edited by the researcher and it was for this reason that there was no request for more news bulletins because it seemed the journalist, though a Sesotho news reader, was not familiar with the writing conventions of Sesotho. The bible chapter as well as the hymnals were also typed in manually. Only one Chapter was sampled regardless of the bible being a large document because the content of the bible mostly involves proper names, most of which are borrowed names as the text is translated and so the names of people or biblical locations were removed from the wordlist.

A corpus is made up of two parts: a header and a body as already mentioned in Section 3.2. A header provides bibliographic or encyclopaedic information of the text, and the body carries the contents of the text. Each text must be provided with its encyclopaedic information for both identification and storage. As such, a header was created for each sampled text using the *TEI-Lite* Header descriptions in order to classify the texts into their domains and genres. Creating headers assisted with determining the different periods, locations, the age groups and gender of the various authors and moreover, the sources/origin of the texts as shown in Table 4.2.

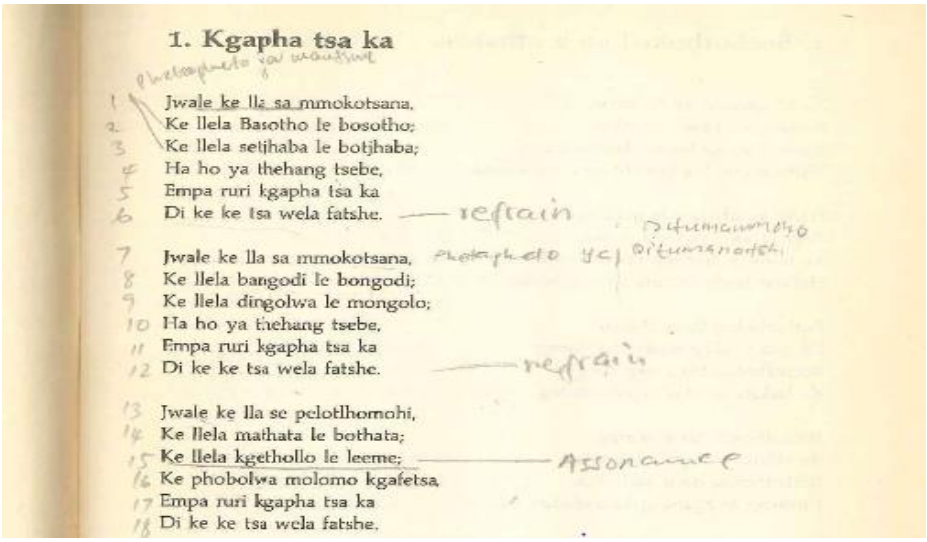
Sample 1.1. Politics: Lesotho Parliamentary Speeches	
Mode	<Written-to-be-read>
Text Origin	<Lesotho Parliament>
Participation	<Multiple Delegates>
Constitution	<Archived>
Preparedness	<Prepared>
Medium	<Speeches>
Style	<Prose>
Genre	<Regulation>
Factuality	<Faction>
Setting	<Formal, Parliament setting>
Function	<National Address>
Topic	<Related to national government>
Technicality	<Semi-technical>
Date	<Years 2006-2012>
Text status	<Published online>
Language	<Sesotho>
Language Status	<Standardized>
Authorship	<Various political delegates>
Sex of Author	<Males>
Age of Author	<35-50>
Region of Author	<Lesotho>
Nationality of Author	<Basotho of Lesotho>
Authority of Author	<Statesmen>
Age of Intended Readership	<All Ages>
Size of Intended Readership	<National Size>
Familiarity of Intended Readership	<Patriots>

Table 4.2: *TEI-LITE Header*

Header descriptions convey critical factors that are of assistance when analysing the differences in spelling and orthography that exist in the corpus; this indicates that a corpus can also be used to assess peculiarities in the different Sesotho writing periods.

Building a corpus involves searching the archives depending on how far back the researcher needs to mine the texts required and this leads to different treatments of the texts, especially, in terms of their age, hence the different strategies of collection as mentioned earlier.

In the case of texts that were scanned, below is an example of the quality of such texts, and the challenges that were encountered from the scanning and conversion processes follows.



1. Kgapha tsa ka

Jwale ke lla sa mmokotsana,
Ke llela Basotho le bosotho;
Ke llela setjhaba le botjhaba;
Ha ho ya thehang tsebe,
Empa ruri kgapha tsa ka
Di ke ke tsa wela fatshe.

Jwale ke lla sa mmokotsana,
Ke llela bangodi le bongodi;
Ke llela dingolwa le mongolo;
Ha ho ya thehang tsebe,
Empa ruri kgapha tsa ka
Di ke ke tsa wela fatshe.

Jwale ke lla se pelotlhomohi,
Ke llela mathata le bothata;
Ke llela kgethollo le leeme;
Ke phobolwa molomo kgafetsa,
Empa ruri kgapha tsa ka
Di ke ke tsa wela fatshe.

Figure 4.2: *Scanned text sample from Kgapha tsa ka (Maphalla: 1993)*

The text in Figure 4.2. is a poem from a prescribed school poetry anthology and a few more poems were taken from this collection, and fortunately, this anthology was also available at the university library and this page was scanned from a copy found at the library. It should be mentioned that with books as old as this one, it is rare to find a copy without handwritten notes and the libraries do not have them stocked in large amounts. This page was scanned as is and further converted into *Word* format using the *OCR* online conversion tool. Figure 4.3. is the result of the *OCR* conversion of the poem:

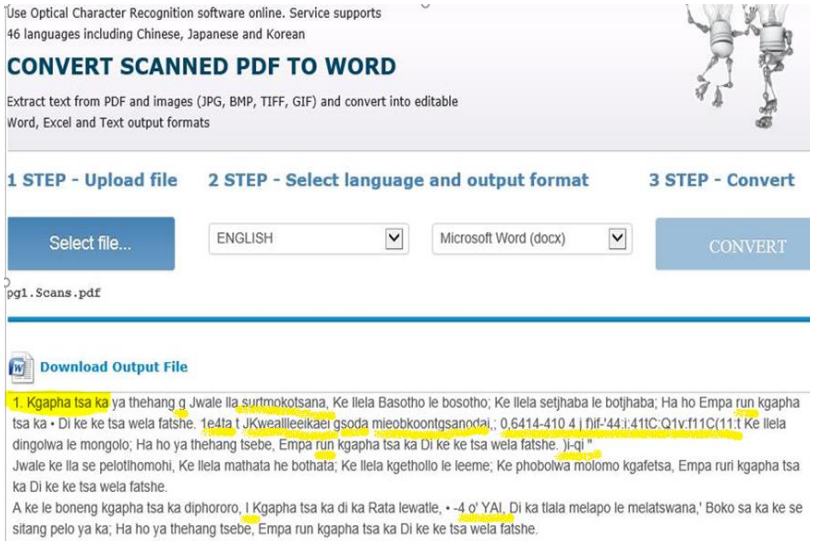


Figure 4.3: *Auto-converted version of the Poem in Figure 4.2.*

Although, overall, the *OCR* conversion tool performed well however, it did not recognize some letters and lines in some cases, as in line 1 and 2 of stanza two where the output is gibberish (highlighted in yellow in Figure 4.3), but it also changed the format of the poem to prose. The sequence of words and their structure is not what we see in the original text in Figure 4.2. This error in processing required post-editing for all the texts that were scanned and processed using the *OCR* conversion tool. In cases where there were many scanning errors, it was decided to retype the document to save time spent on extensive post-editing. The converted text in 4.2. was edited and structured in poetry format as displayed in Figure 4.4.

Jwale ke lla sa mmokotsana,
Ke llela Basotho le bosotho;
Ke llela setjhaba le botjhaba;

Ha ho ya thehang tsebe,
Empa ruri kgapha tsa ka
Di ke ke tsa wela fatshe.
Jwale ke lla sa mmokotsana,
Ke llela bangodi le bongodi
Ke llela dingolwa le mongolo
Ha ho ya thehang tsebe.
Empa ruri kgapha tsa ka,
Di ke ke tsa wela fatshe.

Jwale ke lla se pelotlhomohi,
Ke llela mathata le bothata
Ke llela kgethollo le leeme;

Ke phobolwa molomo kgafetsa, Empa kgapha tsa ka Di keke tsa wela fatshe.

Figure 4.4: *Edited version of 4.3.*

An example of another text that had to be manually typed is the Sesotho newspaper, *Mochochonono* because the images of the articles that were captured from the newspaper were illegible. The articles captured using a camera because the library warned that paper is fragile and scanning would put it at risk of getting torn, and this process was done from the basement area of the Wits University William Cullen Library which did not have good lighting. Figure 4.5. below is one of the images of the newspaper that were captured with a digital camera.

MOCHOCHONONO
THE COMET
Tsa Bergville
Tsa Maseru
FAMOUS PIANO PLAYER AND RADIO STAR
JACOB MOEKETSI
"Do you want to be in Fashion
Smoke Milder, Cooler, Better
Be in Fashion—
Smoke C to C!

Figure 4.5: *Sample of Mochochonono newspaper*

Although it is a Sesotho newspaper, which was distributed in Lesotho and written in Lesotho orthography, it also incorporates other languages for advertisements and some public announcements, therefore, reducing the amount of Sesotho data that could have been sampled

from the newspaper. Nonetheless, even though typing out the newspaper presented more work than scanning and converting would have, the advantage on the one hand was that the English writing found in this newspaper could be omitted from the onset. The newspaper such as the article in 4.5. above, were typed out word for word, the typed version is presented in Figure 4.6. below.

Tsa Maseru

Letsatsi le hlile le fihla moo le hakalang ka mocheso oa lona, hoo ka linako tse ling ho bantehang meriti. Oho ruri u ka haula ba mono ba noang hoenene—u tla fumana joale motho oa bona a felile, 'me joale letsatsi le mo kene le lona.

Har'a beke ha ho ka hoa ba leletho ha ese feela bioscope qetellong ya beke; le ka Moqebelo papali li ne li le siyo. Ho no ho mametsoe lipuo feela tsa hore ba bang ha re Bantu e tla hlola bang ba re Matjantja a tla hlola. Khang ena e le hobane lithimi tsena tse bedi li ne li ts'oanetse ho tsoha li bapala *semi-final* ka Sondaha lebaleng Matlama F. C. mona Maseru. Majantja ea lala e fihlile ka Moqebelo; athe Bantu eona ea tsoha e fihla ka Sondaha.

Ao, ngoaneso, rona bo.montoe-siba-le-kae re se re tseba ho bua hobane ha u n'u le pela batho ba Majantja ua bua ka hare? Bantu e tla hlola, kapa ua bua ka Majantja, u no u tebeloa ka potlako, ebile u bolelloa bore u tla shapuoa. Ka ha nako esa hane ho tsamaya, ea fihla nako ea ho qala ha kena bashemane ha tsoelane pelo hose ngoan'a motle, oa kha-bola—oa qaleha mosebctsi ka lebaleng—papali ea qala ka sekhahla se makatsang, ea ea holimo le tlase, ea re ha phomolo e fihla Majantja ea be e hlabile 1-0. Ha e boetse e qala hape ho ququtana ea thesa ea nyoloha.

Athe batho ha re tsebe hore nyeoe e se e le teng ea Majantja ba tseka hore moteaphala o ile a ja nako hobane pele ho phomolo o ile a ja metsotoso e mehlano. Ha batsamai ba lithimi tse peli tsena ba bitsitseo ke ena hore ba dumellane a tla oketse nako ka lekhetlo la bobeli oa Majantja a ile a hana. Joale ka mora pabali ra utlwa hore nyeo ile ya kena ka baka la pelaelo ea hore moletsaphala o koetse papali ho sa setse metsotso e 15.

Kahlolo ea bolela bore papali ena e ts'oaneloa ke hore e lo phetoa Mohale's Hoek ka Laboraro. Taba e bileng bohloko ho Makaota ke hore ebe mongoli le motsamaisi oa bona ebe ha a tsebisoa. Hape re utloa hore raporotong ea moteaphala ho ngotsoe bore Bantu e hlotse ka 3-2. Na ha se mohlolo hore ba kentse bolo e le ngoe empa ebe e bitsoa tse peli? Joale taba e makatsang haholo ke ea komiti ea B. S. A. ke hore ke eona e laetseng hore lithimi tsena tse peli li tlo bapala moo e tlang ho bona papali empa e re ha papali e felile ebe e amohela lengolo athe joaleka ha e mong oa eona komiti ena, empa e seng a le mong, e le kaofela ha bona ba boneng…

Figure 4.6: *Typed Article from Mochochonono*

Mochochonono is one of the texts in Lesotho spelling, and at the stage of compiling the corpus, the spelling was not reduced to the South African spelling. When it came to lemmas, Lesotho-spelt words were converted to the South African spelling. Some texts featured words from other languages, but the foreign words were not immediately removed from the corpus. The Sesotho articles were retyped as there were, even instances with typos.

The Government Speeches contained most of the code mixing as reflected in Figure 4.7. below where the author of the text gave translations of what he meant in brackets. For instance, ‘results based’ was used to indicate what the author means by *likatleho tse fihlelletsweng*, ‘220, 000 metric tons’ is translated to *lithane tse 220, 000*, while ‘the poverty reduction strategy’ is *Moralo oa Phokotso ea Bofuma*. The translation of concepts within a Sesotho text clarified what the author intended. In cases like this, where it seems like the author was translating what they had already stated in Sesotho, the English was omitted because it did not impact on the full meaning of the sentences even when extracting usage examples from the corpus.

> Letla hopola hore Moralo oa naha oa HIV le AIDS o ile oa hlahlojoa bocha, ‘me moralo oo o seo le teng, ‘me o itšetlehile holim’a tšebetso e bontšang *likatleho tse fihletsoeng* **(results based)**. Lipalo tsa likhakanyo tse matsohong a ‘Muso tsa selemo sena li supa hore khaello ea lijo-thollo ke *lithane tse 220,000* **(220,000 metric tons)**... Tsona ke tse kang: *Moralo oa Phokotso ea Bofuma* **(The Poverty Reduction Strategy)** hammoho le *moralo oa ntlafatso le tlhopho-bocha ea bosebeletsi ba se*chaba **(Public Sector Improvement and Reform Programme)**.

Figure 4.7: *Example of code-mixing in Sesotho texts*

As mentioned earlier that some hardcopy texts had to be typed because the *OCR* was not processing the scanned documents accurately, Figure 4.8. below is one such sample that was typed from a book of hymns. A total of 15 hymns were typed and during labelling, hymns were labelled Genre 4, thus the numbering in the title of the hymn below. In the text below, the hymn was typed as it appears from the original text. Lines 4 and 15 contain words that have diacritics due to contraction meant for poetic effect. What was observed is even though the hymn book was published in South African spelling, there were instances where diacritics featured in the hymnal.

4.9 Seteng Sediba
1. Seteng sediba sa madi
Aletareng ya tefelo
Diba se e leng sehlare
Matl'a sona ke bophelo

2. Baetsadibe ba batsho
Ba se kenang ka tumelo
Ba tloha teng ka basweu
Ka thabo le ka tshwarelo

3. E sa le ke ihlatswa teng
Ke ntse ke bina topollo
Ke be ke kene Moreneng
Madulong a dinyakallo

4. Mohlang oo, ke tla kopana
Kganyeng le bohle ba hlotseng
Ba binang **pel'a** konyana
Rato le ba lopolotseng

5. Ba re amen Halleluyah
Ho boraro bo teroneng
Ntate le mora le moya
Ba bine ka ho sa feleng. Amen

Figure 4.8: *Manually typed hymn sample*

After scanning and converting; typing in; copy and pasting, all samples were compiled into a single *Word* document, a requirement for concordances including *tlcorpus* in order to generate a wordlist. Following which, the unannotated Raw Corpus of 101 359 words was uploaded onto *tlCorpus* and below is what the *tlCorpus* interface looks like:

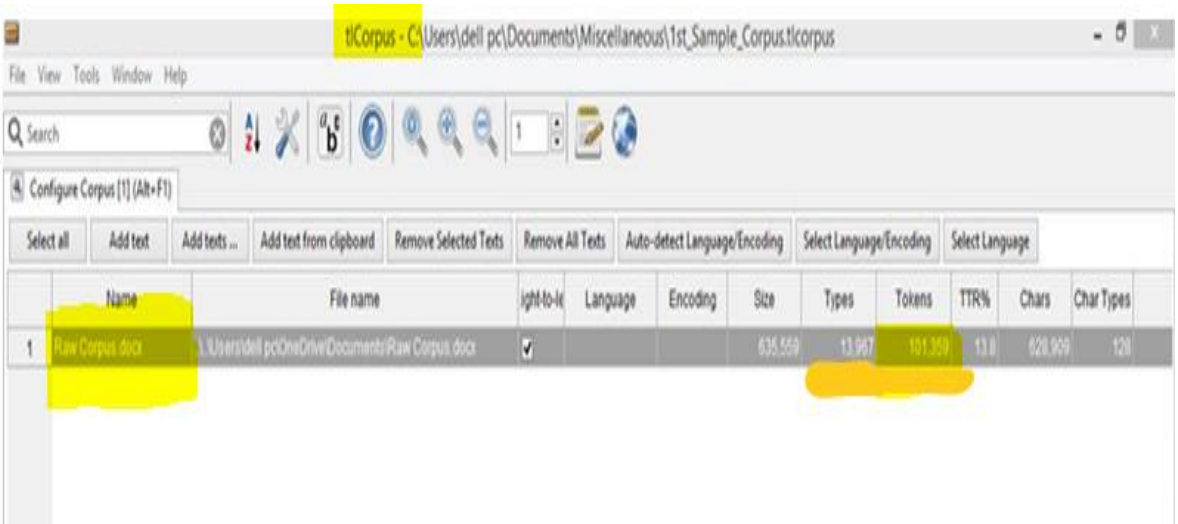


Figure 4.9: *tlCorpus interface for uploading a corpus for Wordlist generation*

The above illustration of the *tlCorpus* interface shows the exact number of 'tokens' in the raw corpus, which is 101 359 although the word 'types' are themselves 13 967 in total, it is the combination of all the text samples in their raw format i.e., without header and body annotations. Turner (2015) explains that "The term "token" refers to the total number of words in a text, corpus etc, regardless of how often they are repeated. The term 'type' on the other hand refers to the number of distinct words in a text, corpus etc." For instance, the word *bana* 'children' may appear a hundred times in a text, and in that case each occurrence is a 'token' whereas 'type' considers the 100 occurrences as one type. After uploading the raw corpus onto *tlCorpus*, a wordlist was generated and it is presented in Section 4.2., under the frequency counts. A decision was made to only take 250 lemmas for the dictionary at this stage, from letters A-B but also ensuring that both major, minor, and hybrid POS are represented.

4.2 Customizing *TLex* for *Sehlalosi Sa Sesotho sa Baithuti*

The title of the prototype dictionary is *Sehlalosi sa Sesotho sa Baithuti* (Annexure 1). The word *sehlalosi* means 'that which defines' and it, therefore, correlates with the purpose of the monolingual dictionary. As the title of the dictionary suggests, this is a dictionary for Sesotho Learners, however, it can be used by any other Sesotho speaker as well. The title was decided upon by doing analyses of the different names that have previously been assigned to Sesotho dictionaries listed in Section 1.4. It is a general dictionary, although, since some of the corpus was sampled from school activity books, medical surveys and news, which are fields that use specialised and technical terms, the dictionary thus contains certain field specific terms.

Figure 4.2.1. below illustrates the properties of the dictionary; it is a Sesotho monolingual dictionary, and its title is *Sehlalosi sa Sesotho sa Baithuti* "Sesotho Learner's dictionary".

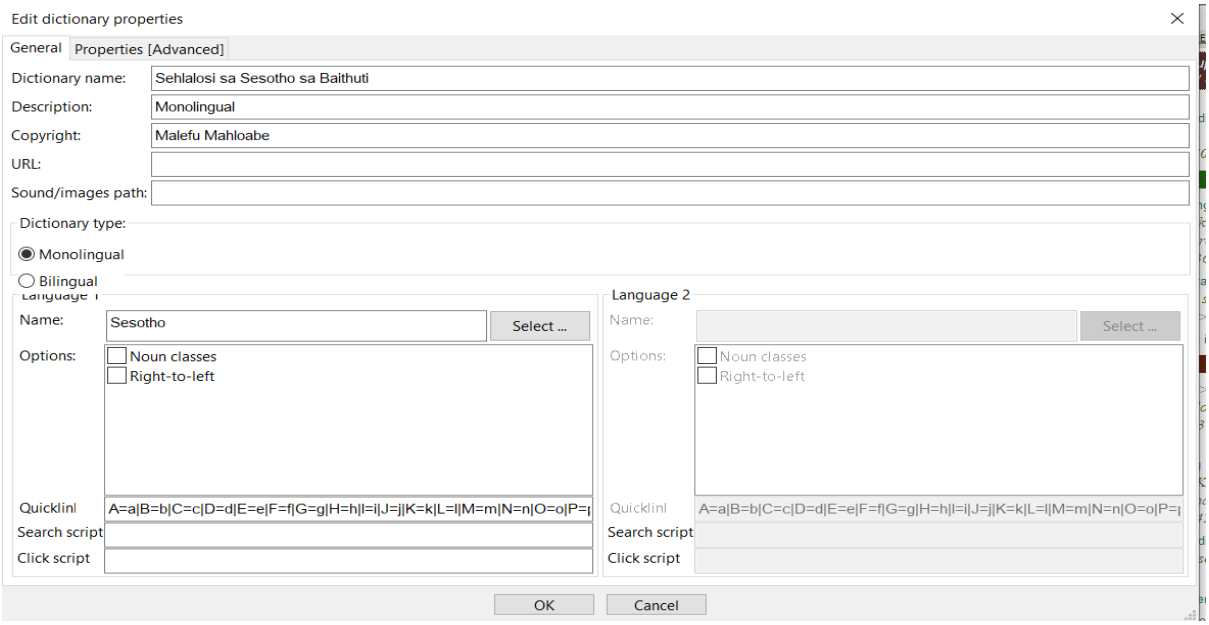


Figure 4.2.1: *TLex interface for insetting dictionary properties*

In 4.2.1., under the **'General'** tab, there is the **'Dictionary name'**, **'Description'** and **'Copyright'**, which are populated. The **'URL'** and **'Sound/images path'** are not yet populated. At this stage, the dictionary is still electronic but not online because the URL has not been created as yet. The sound and images are also not inserted because it is not a priority at this point.

The next steps after creating an identity for the dictionary was to change default mark-up language to Sesotho, which started with renaming POS from English to Sesotho as in 4.2.2. below.

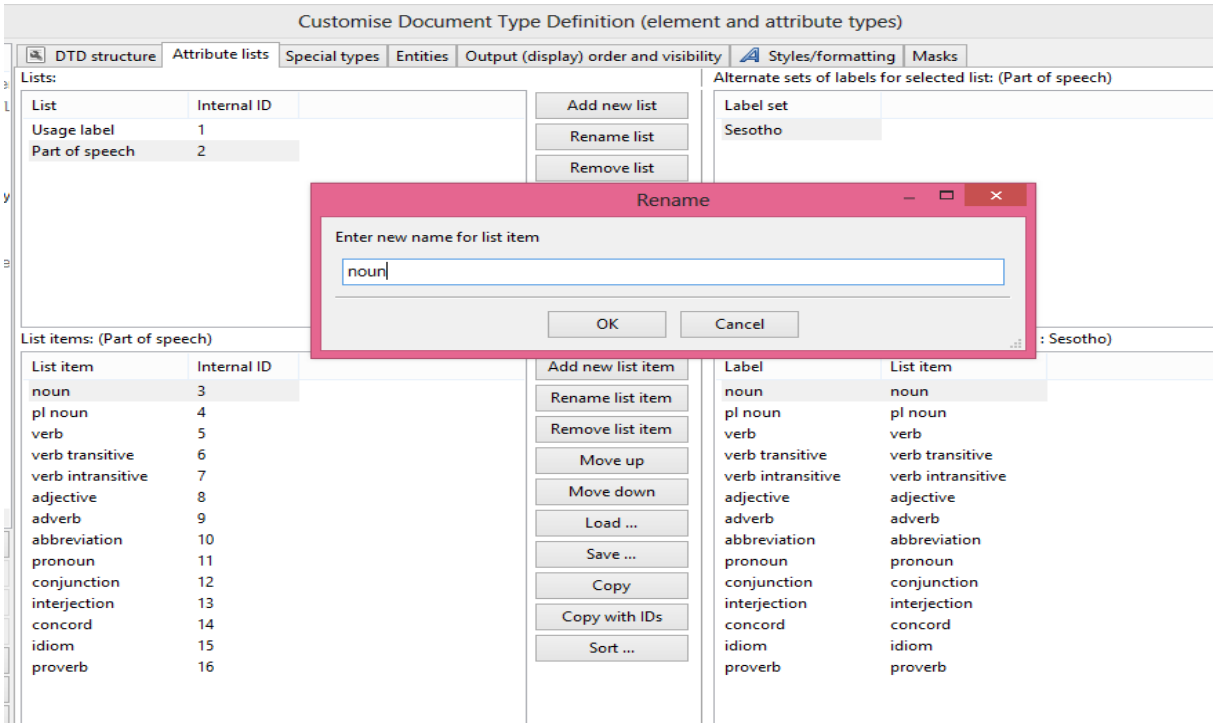


Figure 4.2.2: *Replace English POS names with Sesotho ones*

Following the renaming and replacing done in 4.2.2., the POS were changed as indicated in 4.3.2., also showing the identity number of each POS as this is important for the creator when balancing distribution of POS in the end stage of marking up.

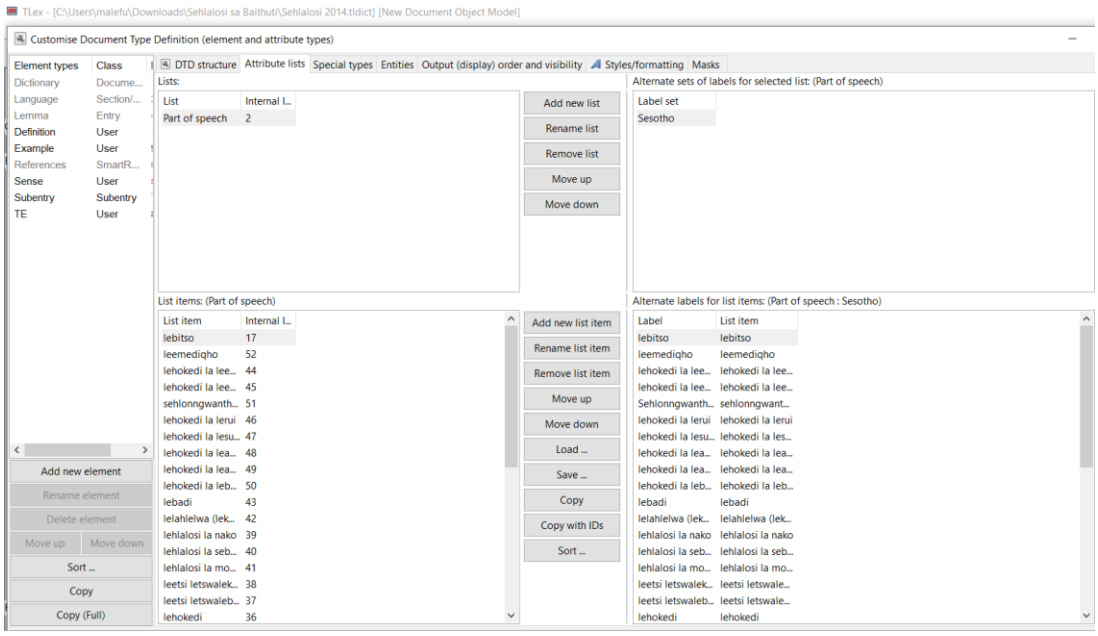


Figure 4.2.3: *POS in Sesotho now as the default language*

Table 4.2.1. below contains the Sesotho POS with their glosses. Some of these POS have subcategories, but in the Table below only the main POS are listed. In Figure 4.2.3. above, the subcategories are inserted for example, the POS *lehlalosi* subcategorises into *lehlalosi la nako* 'adverb of time' (39), *lehlalosi la sebaka* 'adverb of place' (40), *lehlalosi la mokgwa* 'adverb of manner' (41).

POS of Lexemes	**Gloss**
Lebitso	Noun
Leetsi	Verb
Lehlalosi	Adverb
Lekgethi	Adjective
Lesupi	Demonstrative
Lebadi	Enumerative
Leakaretsi	Qualificative
Leamanyi	Relative
Lerui	Possessive
Leemedi	Pronoun
Lelahlelwa	Interjective
Leetsisamodumo	Ideophone
Lekopanyi	Conjunctive
Lehokedi	Agreement markers concords

Table 4.2.1: *Sesotho POS names and their glosses*

After configuring the POS, the other labels that had to be changed to Sesotho where cross-reference types. Since there was no monolingual dictionary to refer to for the equivalents of the cross-references: SYNONYM, SEE, EQUIVALENT, COMPARE, ANTONYM, the English – Sesotho Dictionary (ed. Motsapi, 2015) was used to find translations of these cross-reference types as illustrated in Figure 4.2.4.

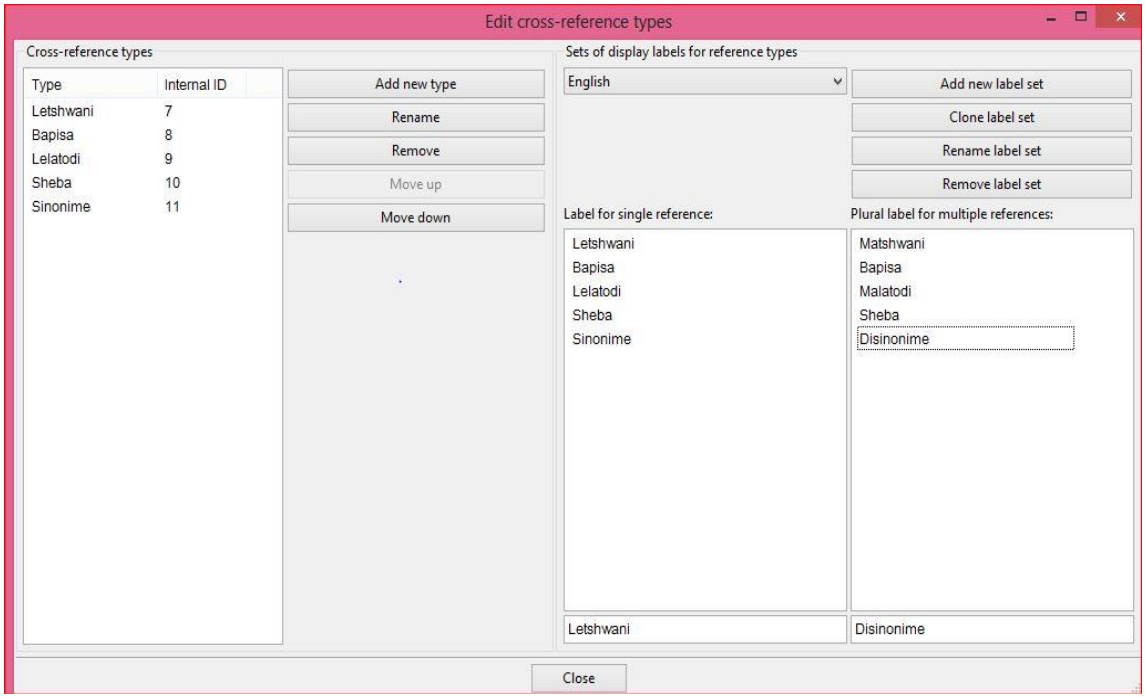


Figure 4.2.4: *Names of cross-reference types changed to Sesotho*

As shown above, the cross-references were changed to:

- *letshwani* "synonym"
- *lelatodi* "antonym"
- *sheba* "see"
- *letshwani* "equivalent (-ce)"
- *bapisa* "compare"

In Motsapi's (2013) dictionary, the "synonym" has 2 more synonyms in Sesotho other than the one *letshwani*, and these are *lehlalosonngwe* and *sinonime*. The 'antonym' also has an additional synonym which is *lehanyetsi* however, both *letshwani* and *lelatodi* were selected because they are common in Sesotho curriculum. While identifying these equivalents, it transpired that the translations for 'synonym' and 'equivalent' are both *letshwani*. Therefore, to move forward, a close look was made into both words to establish why Sesotho uses the same translation for both cross-references. The purpose of pursuing the query into these words was also to determine whether it would be necessary to include both types in the dictionary and name 'synonym' either *lehlalosonngwe* or *sinonime*, although they are less familiar.

Subsequently, *letshwani* was assigned to both because of its usability and understanding that there is no difference between the two.

4.3 Dictionary analysis

Having processed the corpus and generating a wordlist, the data arising therefrom was analysed into dictionary format in order to answer the overarching aim of the research which is to develop a corpus-based monolingual dictionary for Sesotho Learners. The body of the dictionary consists of two major parts which are the macrostructure and the microstructure. The analysis of the body of the dictionary follows below:

4.3.1. Macrostructure

Čechová (2015:13) states that, "The macrostructure of a dictionary is constituted by a general order of entries, that is, of words [which is lexicography are called lexemes or lemmas], their parts or word groups. In a wider sense of the word, macrostructure includes also introduction before the body of the dictionary, tables and appendices." In this case, the narrower meaning of 'macrostructure' is used. After developing a wordlist from the 101 359 words, the wordlist reduced to 13 967 word types and 5464 word types were selected as lemmas or headwords for the macrostructure. All 5464 words were fed into *TLex* in the dictionary interface and the software arranged the entries in an alphabetic order automatically, as reflected in the top-right of the Figure 4.3.1., which shows further properties of the dictionary; from the top on the left, it identifies that the dictionary is written in Sesotho using a tag name ***'Seso'***, which it generated automatically after inserting the details in Figure 4.2. in the previous section. Below the language of the dictionary, is the number of lemmas, the date of the most recent update of the dictionary, and the storage information.

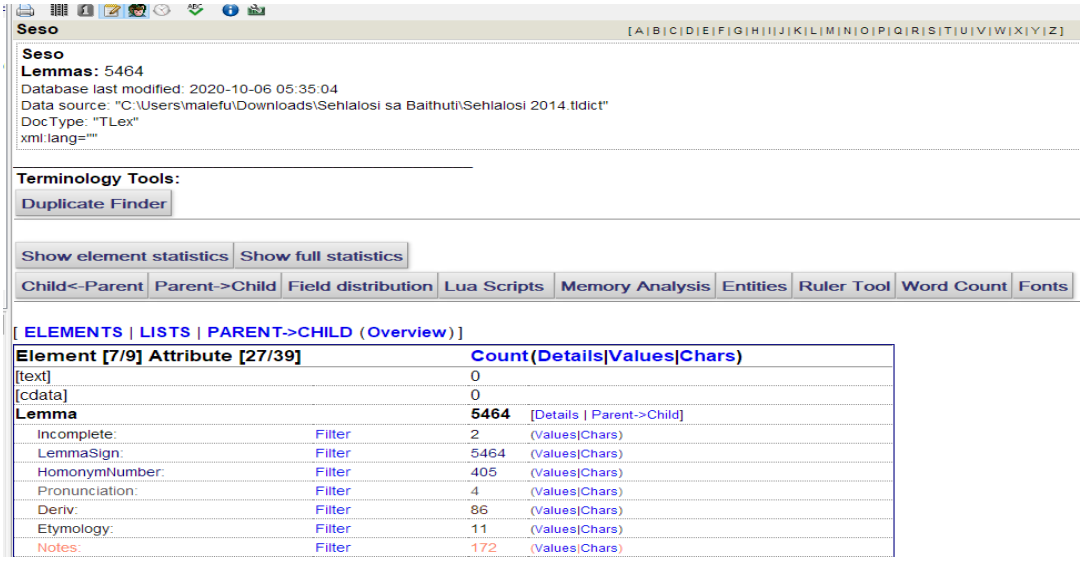


Figure 4.3.1: *Macrostructure of Sehlalosi sa Baithuti*

Additionally, 4.3.1. shows the number of articles that were not yet completed i.e., those that were not yet saved when the developer was making entries at the time. There is also the number of homonyms as detected by the tool, number of pronunciation mark-ups that had not been completed as well as the number of lemmas that were marked as derived, those that had etymological information and those with additional notes. The *'Notes'* bar allows a developer to insert any other information about the lemma or entire article.

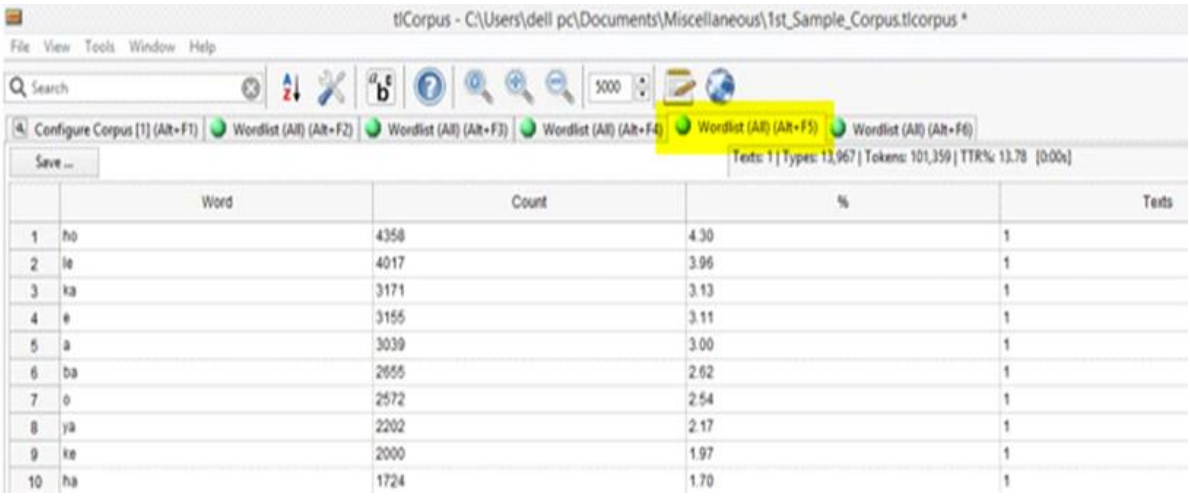
Orthography

Although a portion of the corpus was from Lesotho produced texts (Section 3.2.3), the orthography used in the macrostructure of the dictionary is South African. Words that were found in Lesotho orthography were reduced to the South African spelling once the wordlist was compiled.

Frequency Counts

The wordlist in Figure 4.3.2. below was generated using *tlCorpus* and then exported onto *TLex*. These are just a few words from the wordlist of the 13 967 word types. The highest frequency

words are agreement markers, pronouns, some conjunctives and most interjections and ideophones. However, most of these highest frequency words will be found under letters that are not included in the prototype in Annexure 1 because only ***a*** and ***b*** headwords are included here. Therefore, from Figure 4.3.2. below only the fifth and sixth highest frequency concord are included in Annexure 1. Multiword expressions as complex lexical units were not identified because the concordance tool processes individual word units as can be seen in 4.3.2. and 4.3.3.



	Word	Count	%	Texts
1	ho	4358	4.30	1
2	le	4017	3.96	1
3	ka	3171	3.13	1
4	e	3155	3.11	1
5	a	3039	3.00	1
6	ba	2655	2.62	1
7	o	2572	2.54	1
8	ya	2202	2.17	1
9	ke	2000	1.97	1
10	ha	1724	1.70	1

Figure 4.3.2: *tlCorpus wordlist – high frequency words*

In Figure 4.3.2. above are the following minor classes and agreement markers in the same numbering or sequence in the figure:

(1) ***ho*** = infinitive prefix
(2) ***le*** = conjunction
(3) ***ka*** = instrumental adverbial prefix and manner adverbial prefix respectively.
(4) ***e*** = subject-verb agreement morpheme (also known as subject concord) and object-verb-agreement morpheme (also known as object concord) of Class 9 and both abbreviated as SVA and OVA henceforth.
(5) ***a*** = subject-verb agreement morpheme (also known as Subject concord) and object-verb-agreement morpheme (also known as Object Concord) of Class 6 and both abbreviated as SVA and OVA henceforth.
(6) ***ba*** = subject-verb agreement morpheme (also known as Subject concord) and object-verb-agreement morpheme (also known as Object Concord) of Class 2 and both abbreviated as SVA and OVA henceforth.
(7) ***o*** = subject-verb agreement morpheme (also known as Subject concord) and object-verb-agreement morpheme (also known as Object Concord) of Class 6 and both abbreviated as SVA and OVA henceforth.
(8) ***ya*** = possessive morpheme class 4 and 9.

(9) ***ke*** = copulative formative
(10) ***ha*** **=** instrumental adverbial prefix and manner adverbial prefix respectively.

The ten agreement markers listed above have the highest frequency count and only /*a*/ and /*ba*/ are included in Annexure 1 which contains only 250 lemmas under letters *a* – *b* of the +5000 of the envisaged dictionary.

The middle frequency words are the ones reflected in 4.3.3. and these comprise of nouns and verbs.

2751	bahodisi	2
2752	saluti	2
2753	salesedi	2
2754	sale	2
2755	sarelwa	2
2756	sana	2
2757	sammanamana	2
2758	sajene	2
2759	rupelletsweng	2
2760	rupela	2
2761	rotswana2	2

Figure 4.3.3: *Mid-frequency words*

On the other hand, Figure 4.3.4. is for the words with the lowest frequency as generated using the *tlCorpus* wordlist. The wordlist generated 5503 items and among these were numbers, foreign words and words that were similar but only differed in spelling because of the two orthographies. These were only removed when the wordlist was transferred to *TLex* leaving the total of 5464 lemmas (c.f. 4.3.1).

5490	mophatlalatsi	1
5491	phallelwe	1
5492	meifo	1
5493	diprovense	1
5494	mehwabadi2	1
5495	diqeto	1
5496	mehwabadi1	1
5497	dipuong	1
5498	maholosiane1	1
5499	dipina	1
5500	dipilisi	1
5501	dipontsho	1
5502	diprojeke	1
5503	phanyeho	1

Figure 4.3.4: *Low frequency words*

The low frequency words also belong to the major classes: nouns, verbs and adverbs. Overall, the word types fall under the Sesotho major and minor POS and here, the total of words per POS is only for the 250 sampled for presentation in this research (Annexure 1) and Figure 4.3.5. shows how many words fall under each POS in the 250 sample.

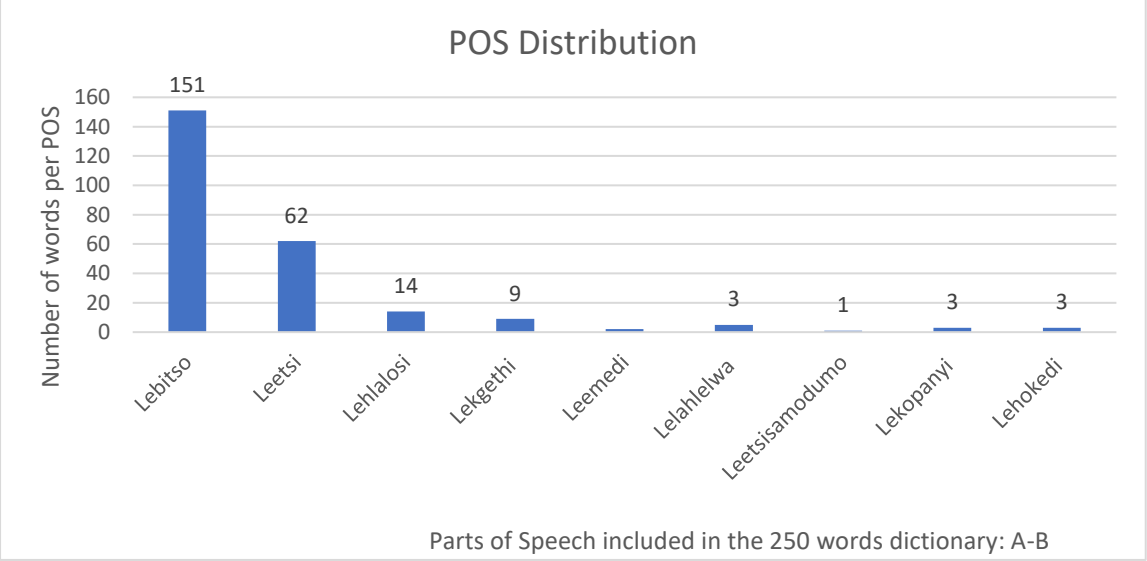


Figure 4.3.5: *POS distribution in the prototype dictionary i.e., Annexure 1*

As reflected above, the noun has the highest number of words in the 250 followed by the verb, pronoun and adverbs so far whereas, conjunctives, interjections and agreement concords are very few. Even though concords make up the highest frequency in the general dictionary, very few appear under A-B letters.

Lexical items

Some of the lemmas in the dictionary are in their primitive form whilst others are in their derived forms and derived forms have full lemmatic status like the default word forms. They were kept in the form in which they appear in the corpus. Further, the lexical items are in both single and plural form and the decision to have a mix of singular and plural was to preserve the words as they were in the corpus because changing the linguistic number to have the singular only would mean that the linguistic number would have to be changed even in the usage examples and that would further require changing other words within the example sentences.

The lemmas include the major and the minor parts of speech. The major parts of speech are nouns, verbs, adjectives and adverbs. Sesotho additionally has the pronouns, conjunctives, ideophones, the interrogative, copulative, possessive, relative, qualificative (see Section 2.2.). Sesotho nouns are divided into classes; the noun has 15 classes which represent singular and plural forms. However, for the dictionary, the lemmas are not annotated with class information because the dictionary is targeted at First language speakers who presumably would be able to distinguish the singular and plural forms. In terms of structure, the nouns are written as full words instead noun stems because that is the form in which they were extracted from the corpus.

POS Mark-up

The POS are written in their full Sesotho names in the dictionary. Abbreviations were considered, but because there isn't an officially standardised and published abbreviation system, this would be problematic.

The macrostructure per POS is made up of the headword, POS and where applicable etymological information as illustrated in the example below:

Example 4.3.1. ***abelwa* 'to be given'**

abelwa: headword

(leetsi): POS

=aba: derivation

Pronunciation

Pronunciation speaks to the phonology and phonetic attributes of words and in dictionary making the pronunciation of words is usually marked using the International Phonetic Alphabet (IPA) system and an alternative approach is syllabification, whereby the syllables making up the word are indicated like in Example 4.3.2. below.

Example 4.3.2. Syllabification of *abelwa*

a-be-lwa

Cross-references, Varieties, homonyms and homophones, multi-POS

The total number of cross-referenced lemmas has not been calculated because not all 5464 words have complete attributes as yet and the same goes for varieties and homophones. However, so far *TLex* has automatically calculated the number of homonyms for the total number of lemmas inserted and this is detailed in Figure 4.3.6. below.

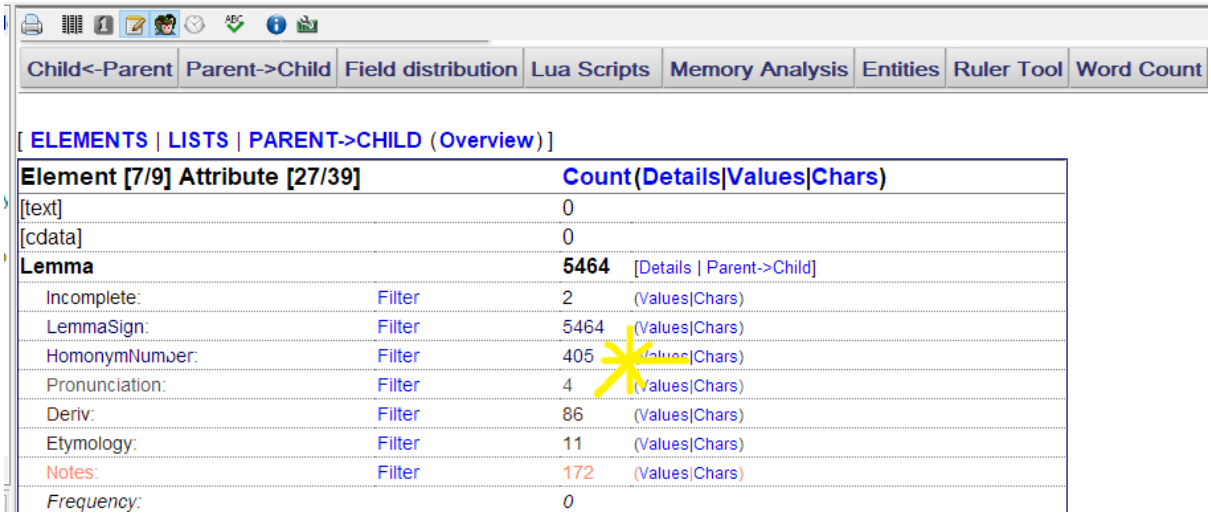


Figure 4.3.6: *TLex pop-up menu with lemma properties statistics, including total number of homonyms in Sehlalosi*

Other than the number of homonyms, Figure 4.3.6. also shows the number of lemmas that had pronunciation mark-up at the time, those that are derivatives, those with etymology and notes elements. The *'Notes'* attribute has been used to provide additional information about some lemmas.

4.3.2. Microstructure of the dictionary

The microstructure of a dictionary is the information associated with each headword. This information is detailed under the headword as an entry. Lehmann (n.d.), states that information belonging to the microstructure includes the following aspects: Entry identity, expression, language variety, structure, meaning, and genetic historical information. In Figure 4.3.7., on the left is the display of the steps of creating entries, which includes selecting the lemma, inserting the sense, definition, pronunciation mark-up etc.

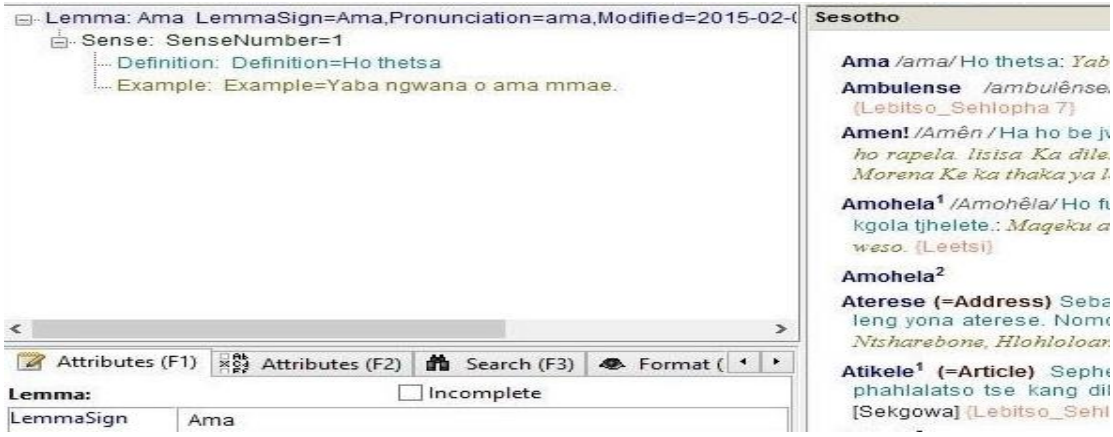


Figure 4.3.7: *TLex Interface for creating entries*

As part of creating entries, usage examples were extracted from *tlCorpus*, and Figure 4.3.8. below illustrates the different contexts from which the word here *ama* 'affect/touch' appear, which were processed using *tlCorpus* and in this way allows the developer to select an example deemed suitable for the purpose of the dictionary. Some example sentences are long, others contain proper names or numbers and therefore, one has to analyse each and decide.

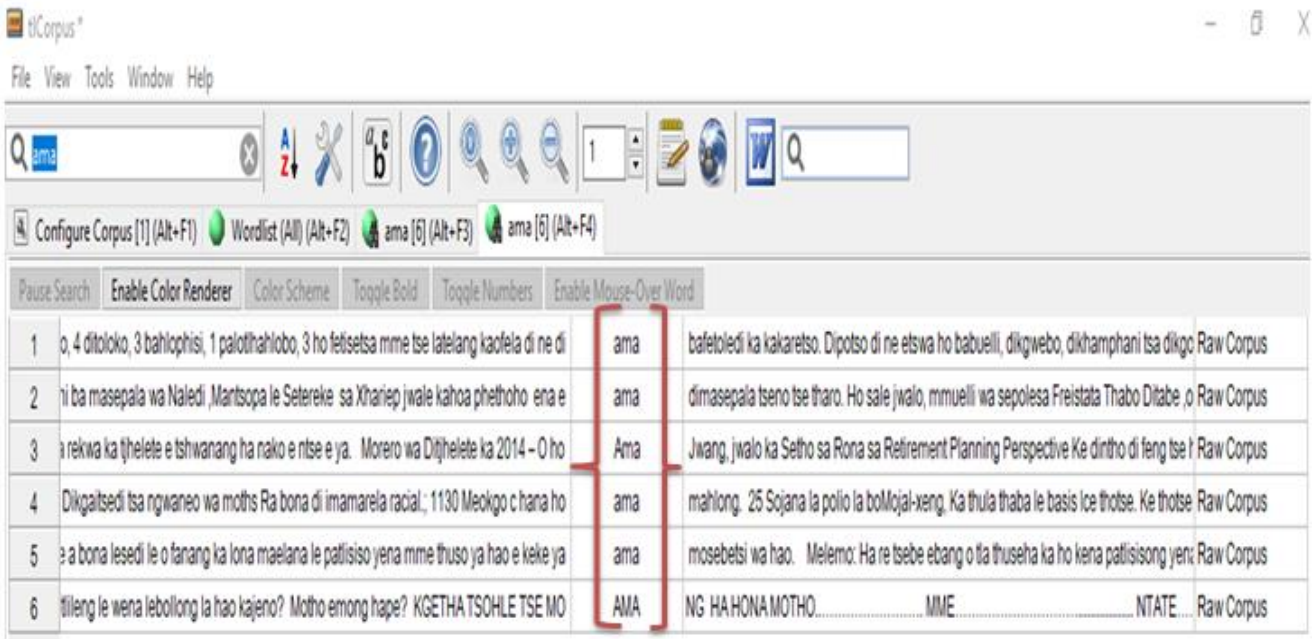
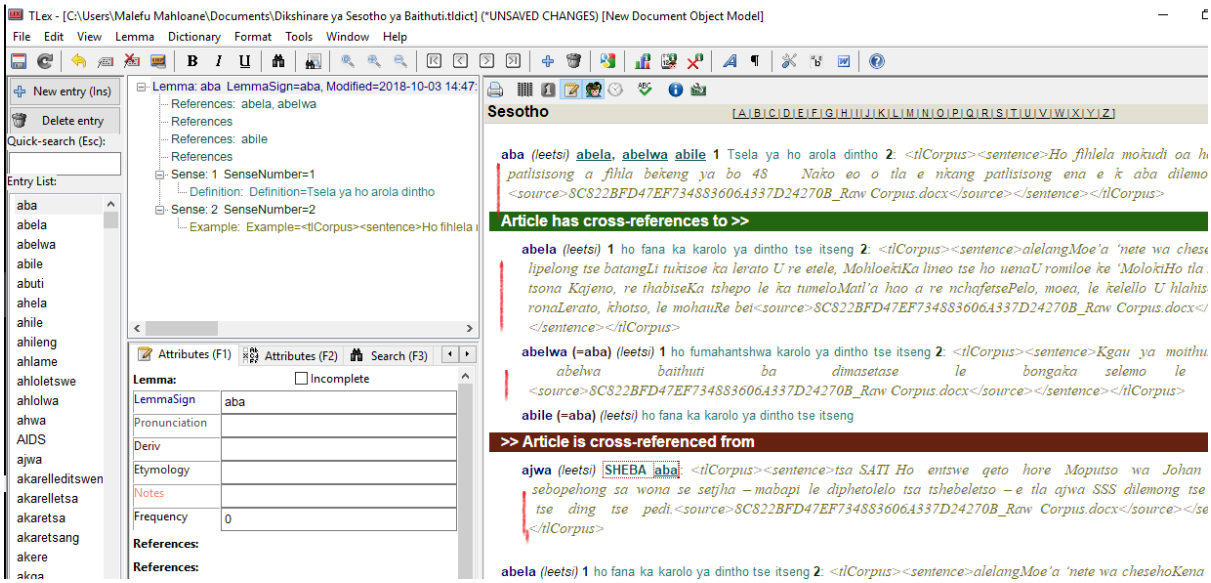


Figure 4.3.8: *tlCorpus concordance interface where usage examples were extracted*

In the Figure 4.3.9. below, the usage examples are added for the word *aba* and its cross-reference types *abela*, *abelwa*, *abile*, *ajwa* and each has a different example, the usage examples are tagged in gold as pencilled in red in 4.3.9. The sentences are tagged in the label *<tlCorpus><sentence> … <source> … </source></sentence></tlCorpus>*, which indicates where the sentence was extracted, and the developer has a choice to edit and remove the source tags and leave only the linguistic information.



4.3.9: *TLex entries interface – Identifying items with cross-references*

The microstructure of the dictionary consists of the internal structure of its entries, that is, it provides linguistic descriptions and, in some cases, encyclopaedic information. This section illustrates how this part of the developed dictionary was populated. The Figure 4.3.2.1 contains the various types of Sesotho nouns that can be found in the dictionary, in the *a – b* batch being submitted with this research. There are other types, which however, appear under other letters, for example, in this sample there are no reduplicated nouns, but those can be found under letter M for instance, and so are included for illustrative purposes in Example 4.3.2.2. below.

Primitive Noun (plural)

bakgotsi */ba-kgo-tsi/* *(lebitso)* batho ba nang le dikamano tsa setswalle: *Bakgotsi ba hao ba tla nahana eng ha o ka bolotswa?*

Proper noun (singular)

Amerika */a-me-ri-ka/* **(=America)** *(lebitsobitso)* dikontinente tse pedi tse nnqa bophirima ya lefatshe, di arotswe ka leboya le borwa: *Boemo ba Loti ho Dolara ea Amerika le bona bo thatafalitse litaba hoba theko ea thepa kantle ho naha, re nkella tsona lijo-thollo, e nyolohile haholo.* [Senyesemane]

Loan noun (singular)

atikele */a-ti-ke-le/* **(=article)** *(lebitso)* Sephepheshana se ngotsweng mabapi le taba e itseng se fumanwang kahara bukana tsa phahlalatso tse kang dikoranta: Ha o batla ho nehelana ka atikele mabapi le kgatiso tse tlang, re tsebise. [Senyesemane]

Common noun (plural)

bakwena */ba-kwe-na/* **(=kwena)** *(lebitso)* batho ba anang kwena: *E pota e kgonya mmadihlofana Ke kgonya ka leqoaphe setholehadi Eitse se kgonya Basotho ra tsoha, Ke kgale le kgotsa ho lapa, Bakwena; Ho mengwaha le supaka ditlhobolo.*

Figure 4.3.2.1: *Mark-up of nouns*

The primitive noun in the above figure, shows syllabification mark-up in slating lines, POS attribute, definition and usage example. Because it is a primitive/primary Sesotho word it does not have the derivation and etymology aspects unlike the proper noun and borrowed noun below it. Looking at the proper noun and loan noun, their source is English which is illustrated in round brackets preceded by an equal sign, and this is important as it shows the relation between the Sesotho adopted spelling and the source language spelling, and also shows how sounds were adapted from one language to the other during the borrowing process. The last

article, the common noun, is different from the three in that it has a derivation attribute, *kwena* ‘crocodile’ but here it is derived from a Sesotho word, and therefore, it was found unnecessary to include its etymology because it is a primary Sesotho word. All proper nouns in the dictionary start with a capital letter to show the distinction from common nouns. Although here, the noun number (singular or plural) is identified in these examples, in the dictionary, this information does not appear yet because a decision is still to be made when expanding the dictionary to either convert all plural nouns to singular as this is a standard tradition in dictionary writing. The decision to leave plural nouns as they are at the moment is based on the background that these words were taken from the corpus as they are, and it helped to keep them that way because this made it easy to extract usage examples from *tlCorpus*.

The noun that follows is a reduplicated one. As mentioned earlier, there were no reduplicated nouns found in the A-B batch and this noun had to be taken from other letters in this dictionary and the same is done throughout this analysis for POS that were not found under A-B, mostly adverbs and qualificatives and the minor classes.

Reduplicated noun

motwaitwai */mo-twai-twai/ (lebitsophethapheto)* **SINONIME** **moleleri** motho ya tsamayang a le mong:

ka tlallwa ke naha se ka motwaitwai, moleleri.

Article has cross-references to >>

moleleri (=Lelera)tlex://ATTR:108.108,7/

Figure 4.3.2.2: *Mark-up of reduplicated noun*

The noun *motwaitwai* ‘solo traveler’ has *twai* as a stem and it repeats itself without repeating the class Prefix. In this article, the synonym is also included as well as its derivation. The

designation for nouns is *lebitso*, but this category can be classified further to differentiate the different kinds/types of nouns and so these names have been used in both 4.3.2. *lebitsobitso* 'proper noun' and *lebitsophethaphetho* 'reduplicated noun' above to capture these distinctions. See complete list of Sesotho categories in Table 4.3.1.

It is common practice for languages to derive new words from primitive forms in order to increase their lexical stock. These derivative processes were outlined in Section 2.2.1. (Morphology) and were analysed in Figure 4.3.2.2. as examples of derived noun types in the dictionary.

Deverbative noun (plural)

baahi */ba-a-hi/* **(=aha)** *(lebitso letswaleetsing)* Batho ba phelang tulong e itseng: *tsejweng a thulwe ke koloi mmileng wa R34 pakeng tsa Welkom le Kroonstad ,mme ho fihlela ha jwale ha ho motho ya se ahlahelletse ka pele ya ka tswang a tseba mosadi eo ,setopo sa hae se bolokilweng moshareng wa mmuso Welkom h otloha esale kotsi e etsahala moqebelo o fetileng 10/05/2013*

Compound noun (plural)

bajalefa */ba-ja-le-fa/* **(=ja + lefa)** *(lebitsikgoboka)* batho ba fumanang le ho sala le mafa akang thepa ya motho ha a hlokahala: *ke tokomane e bohlokwa hoba e tla thusa Balaodi ba Mokga ha ba lefa bajalefa ba hao ditjhelete.* {lebitso lena le bopilwe ka kutu tse pedi, kutu ya leetsi ja le kutu ya lebitso lefa.}

Figure 4.3.2.3: *Mark-up of derived nouns – Part 1*

Adverbial noun (singular)

bosiu */bo-si-u/* *(lebitso letswalehlalosing, lehlalosi la nako)* ha lehodimo le fiifetse letsatsi le sa tjhaba : *boMopheme ba ntlola bosiu.*

Adjectival noun (singular or plural)

bobe */bo-be/* **(=be)** *(lebitso letswalekgething)* ntho e kotsi kapa e tshabehang: *Empa a dumela hore motho le motho o na le botle le bobe esitana kgang e nngwe le e nngwe e na le mahlakore a mabedi.*

Figure 4.3.2.4: *Mark-up of derived nouns – Part 2*

As shown in Figures 4.3.2.3. and 4.3.2.4., all derived nouns have the derivation attribute in addition to the pronunciation, POS, definition and usage mark-up. The root of the deverbative *baahi* 'builders or residents' is the verb *aha* 'build or reside'. The two differences that can be seen in the above figure is that first, when it comes to the POS of the lemma, the specific name for the type of noun – *lebitsokgoboka* and *lebitso letswaleetsing* - is given to assist the learners with the appropriate terms for the different classes of nouns and secondly, an additional description is added at the end of the article. The text in the coral colour is a 'Note'. *TLex* allows the developer to make additional notes to articles and in this instance, the note explains that the compound noun is made up of a verb stem plus a common noun stem.

The adverbial and adjectival nouns are formed with adverbial and adjectival stems. The adverbial noun falls under both the noun and the adverb of time thus the two POS, however, its meaning or usage does not change. On the other hand, the adjectival noun is formed with the adjective /-be/ whose prefix can change depending on the prefix of the noun that it qualifies.

The qualificatives and adverb articles in the dictionary are structured similarly, with the difference being the POS itself. Adverbs are subcategorised according to place, time or manner.

There are several lemmas with cross-references in the dictionary such as the one in 4.3.2.5. below. In cases where a cross-referenced word is a synonym, only the pronunciation, POS and cross-reference type is inserted in article of the synonym, without the definition. For instance, in the example below: *basetsana* is a synonym for *banana* “girls”, and because they are synonyms, there was no need to provide a definition for the latter as it has already been supplied under *banana*, which in terms of aphabetic arrangement, comes before *basetsana*.

Cross referenced item (plural noun)

basetsana */ba-se-tsa-na/ (lebitso)* **SINONIME banana**

Figure 4.3.2.5: *Mark-up of a word with a cross-reference*

As was mentioned in Section 4.2. subject concords had the highest frequency count in the wordlist but here only one is analysed. The articles for concords are also structured in a standard manner as the major POS, they include pronunciation mark-up, label (equivalent of POS), definition and usage example - as illustrated next in 4.3.2.6.

Class Prefix and Concord

a */a/ (Sehlonngwanthao)* **1** sehlongwanthao sa mabitso a sehlopha sa pele: *matjhaba a utlwe Sesotho melodi* **2** *(lehokedi)* lentswe le bontshang kamano dipakeng tsa sehlooho le leetsi polelong: *e se e le moshanyana a tsamayang ntle le ho itshwarella* **3** *(lehokedi la leemediqho)* lentswe le bontshang kamano pakeng tsa sehlooho le leemediqho polelong: *bao matla a puso a dutseng ho bona* **4** *(lehokedi la lerui)* lentswe le bontshang kamano pakeng tsa lebitso le lerui : *le ba tshireletso ka mafapha ohle a bona* **5** *(lehokedi la lesupi)* lentswe le hokelang lebitso ho se etswang: *Mphemphe e a lapisa, Bashweshwe* **6** *(lehokedi la leakaretsi)* lentswe le bontshang kamono pakeng tsa sehlooho le leakaretsi: *matswalla a Lesotho* **7** *(lehokedi la leamanyi)* lentswe le hokelwang pele ho leamanyi: **8** *(lehokedi la lebadi)* lentswe le hokellwang pele ho palo: *wekshopo ena ya mahatsi a mararo e reretswe ho neha bankakarolo dintlhatheo*

Figure 4.3.2.6: *Mark-up of agreement concord*

The concord *a* has various senses and in the above, the senses 1, 2, 3, 4, 5, 6 and 8 have all five standard attributes. Sense 7 does not have a usage example as this was not found in the corpus.

Pronouns

There are 27 pronouns in the batch of 250 words in Annexure 1 and some Sesotho pronouns are ambiguous in that they sometimes fall under another POS - multi POS words - or have a different function altogether in some contexts, and not all of them are discussed in this section because they are annotated in the same pattern as all other words in the dictionary. The absolute pronoun *bona* 'them' or 'this one' in 4.3.2.7. below was selected for analysis particularly because it is also a primitive verb *bona* '*look/see'*. In cases where a word belongs to multiple

POS, the word is ascribed as many senses as are possible and these are marked using numerical numbers as can be seen below.

Pronoun

bona */bo-na/ (leemediqho)* **1** *ba bane bona ba ne ba sa tswa dikotsi* **2** *(leetsi)* ho tjhadimehela ho itseng ka mahlo kapa kutlwisiso: *marena, boradipolotiki, baruti ba moshoeshoe le baphatlalatsi ba e bona*

Figure 4.3.2.7: *Mark-up of lemmas belonging to more than one POS/function*

The properties of the word as a pronoun are provided and as a verb are both included in the article in 4.3.2.7. above. The first sense is a pronoun *'leemediqho'* and the second sense is a verb *'leetsi'*. A single article was created for all homographic and homonymic words as in this example. As illustrated in 2.2.3., pronouns have various types and for the purpose of this analysis only one is included here because their mark-up is the same throughout and the difference is whether the pronoun has other senses in which case it is detailed in each article in Annexure 1.

Qualificatives are another category that bears a very close relationship to the noun as its function is to describe the noun as explained in Section 2.23. Some qualificatives were not found under the letters A-B and have been extracted from other letters for the analysis.

The four types of qualificatives are listed in 4.3.2.8. below:

Qualificatives

Adjective

disele */di-se-le/ (lekgethi)* dintho, tse eseng tsona mme di sa nepahala : *o paka ntho disele mme o tla kena kotsing*

Relative

boima */bo-i-ma/ (leamanyi)* ha tekanyo ya ntho kapa sebopua e le maatla haholo : *o ne a jere thepa e boima bo makatsang*

Enumerative

mong */mo-ng/ (lebadi)* ho ba bonngweng : *mme ke tshepa hore ke tla ba e mong wa sehlopha sa bona sa ditoloko*

Possessive

hao */ha-o/ (lerui)* lentswe le sebediswang ho bontsha hore ntho ke ya motho: *Na kgopolo ya lebollo la banna la sepetle le kgahlano le ditumelo ha hao*

Figure 4.3.2.8: *Mark-up of qualificatives*

There are no peculiarities in the mark-up of these POS as they all have the 5 standard attributes. The adjective *disele* means 'bad ones', the relative *boima* is 'heavy', the enumerative *mong* is 'each or every or one of', and the possessive *hao* is 'yours'.

Similar to qualificatives, the mark-up of verbs follows the same patterns although there are minor differences here and there because of its complex nature i.e., verbs are classified according to tense including immediate and remote tenses; verbs can occur in the positive or negative; active or passive; they use various extensions; can be derived from other word categories and in all cases a verb must always agree with its subject and object as shown in Section 2.2.2. Most of these properties are shown by affixes and it is general practice in lexicography for the verb lemma to be represented by the verb stem only. For example, Figure 4.3.2.9. below illustrates the article of the primitive verb *aba* in the default and past tense forms

and some of its derivatives with verbal extensions such as /-el-/ and /-w-/. All these forms are lemmatised, and the cross-references are used to show their relationship to the primary verb.

Verb (primitive verb)

aba */a-ba/ (leetsi)* **BAPISA abela, abelwa, ajwa abile** ho fana ka dintho :*Puso e tla aba baahi bohle ka matlo.*

Article has cross-references to >>

abela (=aba) (leetsetsi) 1 ho fana ka karolo ya dintho tse itseng ho tla re abela tsona Kajeno, re thabise

abelwa (=aba) *(leetsetswa)* **1** ho fumahantshwa karolo ya dintho tse itseng: *e tla abelwa baithuti ba dimasetase le bongaka*

ajwa (=aba) *(leetsetswa)* **SHEBA aba**: *e tla ajwa SSS dilemong tse ding le tse ding tse pedi*

abile (=aba) *(leetsi)* ho fana ka karolo ya dintho tse itseng

buisana */bu-i-sa-na/* **(=bua)** *(leetsi)* ho tshwara taba pakeng tsa ditho tse pedi kapa ho feta.: *ho buisana ka tshebediso ya diphoofolo ho tsa temo*

Figure 4.3.2.9: *Mark-up of a primitive verb and its associated derivative and word forms*

The primary verb *aba* 'distribute' occurs in 6 different forms as either verb derivatives - the applicative *abela* and passive *ajwa* and *abelwa* or the relative (type of qualificative) as in *abelwang* and *abetsweng* where the latter is also in the past tense in the same way as *abetswe* and *abile* - from the corpus, but only four are included in the dictionary as illustrated in 4.3.2.9. above.

Over and above deriving from itself by means of verbal extensions, the verb also derives from other word categories such as the noun, the adjective and the ideophone as in Figure 4.3.2.10. and the mark-up of these derived forms is similar to that of other lexical categories.

Noun derived verb

hlalefa */hla-le-fa/* **(=bohlale)** *(leetsi letswalebitsong)* ho buleha ba le kutlwisiso e batsi: *mme a lemoha hore ha se ka ho hlalefa ha hae hore ebe o fihlelletse maemo ao*

Adjective derived verb

ngatafala */nga-ta-fa-la/* **(=ngata)** *(leetsi letswalekgething)* ho ata ho feta ka qalong: *ya re ha bohle ba utlwa sello sena, letshwele la ngatafala la bokana lebaleng*

Ideophonic Verb

bjwaramanya */bjwa-ra-ma-nya/* **(=bjwara)** *(leetsisamodumo)* ho dihela kapa ho lahela ntho fatshe kapa moo ho leng thata hore ebe e qetelle e tjhwatlehile : *a fihla a di bjwaramanya majweng.*

Figure 4.3.2.10: *Mark-up of derived verbs*

There are also words in the Sesotho lexicon whose function is to qualify the verb, and these are called adverbs and they subclassify into three main types – manner, time and location. Each type was lemmatised and named accordingly instead of using the umbrella term *lehlalosi* 'adverb' as illustrated in 4.3.2.11. below, and this is to help the user not only to identify the differences.

Adverbs

Adverb of place

basothong */ba-so-thong/* **(=Basotho)** *(lehlalosi la sebaka)* moo Basotho ba bokaneng kapa ba ahileng ka bongata teng: *Phelong ba kajeno, o tsebe ka dibuka, o se ke wa jewa, 45 Lehelo le kola Basothong ba jwale, Ho atlehisa naha*

Adverb of time

bosiu */bo-si-u/ (lehlalosi la nako)* ha lehodimo le fiifetse letsatsi le sa tjhaba : *boMopheme ba ntlola bosiu*

Adverb of manner

hantle */ha-ntle/ (lehlalosi la mokgwa)* ka mokgwa o amohelehileng : *ke hantle ha e le moo a hlokometse tse leng tsa hae*

Figure 4.3.2.11: *Mark-up of different types of adverbs in the dictionary*

Adverbs also double-up as classes 16, 17 and 18 nouns as in the example of the adverb of manner above, which is also a class 16 noun. The locative, *basothong* in the diagram is an adverb of place formed from the noun *Basotho*. This is not the only way in which adverbs are formed from nouns. There are two other ways in which *ho* 'to' or *ha* 'at the place of' is prefixed to the noun and such cases follow the same lemmatisation structure as above.

The following analysis will consider fixed or minor categories such as interjections, ideophones, conjunctives and clitics and the hybrid category which also follow similar mark-up patterns, which include pronunciation; derivation if they are not primary Sesotho words; definitions and usage examples. There are a few interjections and conjunctives in the *a – b* batch. Looking at the interjective in 4.3.2.12. below, it also has an etymology attribute, which is *Seheberu* "Hebrew".

Interjective (borrowed)

Amen! */a-men/* **(=amen)** *(lelahlelwa)* Ha ho be jwalo. Ke lentswe le kwalang thapelo ya Bakreste: *lisisa Ka dilemo-lemo Wa mphumana, wa mphedisa Wena Modimo Chorus Ke phela ka boiketlo atleng tsa Morena Ke ka thaka ya leihlo Le nna ho yena. Amen.* [Seheberu]

hleng */hle-ng/* **(=hle)** *(lelahlelwa (lekgotsi))* lentswe le sebediswang ho ha tella, ho kgotsa, kapa ho bontsha makalo polelong: *Baheso, hleng, elellwang ditsietsi mme le fadimehe*

Conjunctive

athebane */a-the-ba-ne/ (lekopanyi)* **BAPISA athe** lentswe le sebediswang ho atolosa tlhakisetso: *esale eka re latola phetoho; athebane re batla ho itlhonepjhwe*

Ideophone

be */be/ (leetsisamodumo)* lentswe le sebediswang ho fetelletsa ketso kapa tshobotsi : *o ne a tsebahala e le motho a mobe a re be!*

Clitic

hle */hle/ (lelahlelwa)* lentswe le sebetsang ho atella maikutlo : *re kopa hore bobodu bona bo fetiswe hle ka tswetswe*

Figure 4.3.2.12: *Mark-up of minor classes*

The language of origin for the interjective above is written in Sesotho here the same way as it was done for the borrowed nouns in Figure 4.3.2.1. where the language of origin for the words *America* and *Article* is written in the Sesotho equivalent i.e., *Senyesemane* for English. Words like *Senyesemane* and *Seheberu* are acceptable and common translation equivalents for these languages. The conjunctive *athebane* 'whereas' is sometimes written in short as *athe* in the wordlist, which is most popular is spoken language, however, *athe* as a second sense where it is used as an interjective that expresses surprise or disbelief thus, the cross-reference used in this case is *BAPISA* 'COMPARE' instead of SINONIME.

In Section 2.2.6., the ideophone was classified as a hybrid category because verbs do derive from it as in *bjwaratsa* 'break into pieces' from *bjwara* 'sound of breaking' in Figure 4.3.2.9 and the reverse is also true because ideophones derive from the verb. The ideophone in 4.3.2.12. above exaggerates the intensity of the ugliness and the ideophone.

The final fixed category analysed is the clitic, simply explained as a leaning on word. In the main, it is a type of affix that always attaches to a word without changing its category though there are a few cases where it is simply a word form as in *hle* 'please'. Figure 4.3.2.12. treats *hle* 'please' as a clitic and *hleng* is an interjective derived from *hle* by suffixing [-ng] and the difference between the two is indicated in the POS and definition as *lelahlelwalekgotsi* and *leitshetlelli* respectively.

4.4 Results

After having applied the design criteria set out in Chapter 3, the following results were achieved: a raw corpus of 101 359 Sesotho words has been compiled in order to provide the data required for the dictionary, the corpus was used to generate a wordlist that was transferred to *TLex* to create the now prototype Sesotho Monolingual Learner's dictionary. The prototype

contains 250 lemmas with complete articles that consist of definitions, POS, pronunciation mark-up, examples and in some instances etymology details in Sesotho has been developed. These results will be discussed in the next chapter.

4.5 Conclusion

This Chapter presented the processes followed to create the Sesotho monolingual learner's dictionary. It highlighted the formats of the macrostructure and microstructure of *tlCorpus* and *TLex*, the main tools used for corpus processing and dictionary development in this research. Although the aim is to have an online dictionary, at this stage, the dictionary is only electronic primarily because the version of *TLex Suite* used here does not allow permissions for online publication. The developer will have to upgrade the license to achieve this. Nonetheless, the prototype presented is an indication of the headway made towards developing a fully-fledged Sesotho monolingual Learner's dictionary. The next Chapter will further discuss the overal findings of the research.

CHAPTER 5: DISCUSSION OF RESULTS AND CONCLUSION

Introduction

The purpose of this chapter is to discuss the results of this research. This research aimed to develop a monolingual online Sesotho Learner's dictionary using Corpus Linguistics approaches, specifically, the corpus-based approach. The overarching question in this research was: What are the factors of consideration when developing a dictionary with these specifications? This question was further broken down into the following sub-questions: What kind of data and tools can be used to develop a monolingual online learner's dictionary? What kind of information can be marked-up for a monolingual dictionary that can be useful for language learning? Through this research, two main results are firstly, a written, unannotated, multi-genre corpus that comprises of texts written in both RSA and Lesotho orthography was compiled. Secondly, a prototype monolingual Sesotho learner's dictionary which has been given the title *Sehlalosi sa Sesotho sa Baithuti* has been developed and will be expanded with the purpose of making it an online resource. These results will be discussed in relation to the literature reviewed and methodologies used. The rest of the chapter will look at what these results mean; what the study is contributing to Sesotho lexicography going forward; its limitations, and a conclusion which includes future work and recommendations.

5.1. Interpretation of results

This section provides the analysis of the results reported in Chapter 4 during the analysis of the data and each result will be related to the research question posed in Chapter 1.

5.1.1. A response to the first question of this research - *What are the factors of consideration when developing a dictionary with these specifications?* - a wordlist generated from corpus is required to populate the dictionary as the corpus-based approaches are the way to go in current dictionary compilations. The method used for the development of this dictionary is corpus-

driven. The corpus-driven approach is one of the CL approaches. “Corpus-driven research is more inductive so that the linguistic constructs themselves emerge from analysis of a corpus” (Biber, 2010: 206). It concentrates on the “observation of certain patterns that lead to the research hypothesis” (Tognini-Bonelli, 2001: 17). A corpus is a collection of written and/or spoken text samples which reveal the structure and use of a language (McEnery and Wilson, 2001; Sinclair, 1991; Kennedy, 2014), and it can be used as data for any linguistic, sociolinguistics, computational linguistics or lexicography research. It has to be balanced and representative of a language to allow for research that is quantifiable and comparative (Biber et al. 1998). Since the 1970’s an additional defining component of a corpus is that it has to be machine-readable (Kennedy, 2014: 4). In fact, McEnery and Wilson (2001:31) state that the term ‘corpus’ almost always implies the additional feature ‘machine-readable’. Its other characteristics are, therefore, finiteness, balance and representativeness, a variety of samples and serve as a standard reference. A finite corpus is a complete corpus without the possibility of being added upon, for example, a corpus compiler can decide to compile a corpus of one million words and never grow it.

Balance and representativeness on the other hand, are required to ensure that a corpus consists of the various genres that are available in a language. There are at least 29 established text types attributes according to Atkins, Clear and Ostler (1992, pp. 6-9), which encompass many different text genres and yet many of these are not available in Sesotho such as academic, scientific communication and journalism texts to mention a few. In spite of this lack, the corpus compiled for this genre is multi-genre as far as is possible and as a consequence, as representative as reasonable for the purpose of the learner’s dictionary. The corpus developed for the dictionary presented in this research will be enhanced as more genres become available in the foreseeable future to make it more representative. Further, on the factor of representativeness, the inclusion of Lesotho Sesotho texts in the corpus added to the vocabulary

because for example, in South Africa one does not find parliamentary speeches available in Sesotho.

The corpus used in this study is composed of 101 359 orthographic words and was developed by the author at the beginning of this research as there was paucity of Sesotho corpus. It is a monitor corpus i.e.; it is open-ended, and thus more texts will be added over time. It contains two Sesotho varieties - the standard Sesotho of Lesotho and the standard South African Sesotho. There are five texts types in the corpus; from samples of 12 text genres from various periods ranging from the 1830s to 2015, it is diachronic in this manner. In the 1960's, corpora sizes ranged from as little as 50000 words as in the Lancaster/IBM spoken corpus which is 80,000 tokens. The current corpus does not meet the standard, minimal recommended amount of a corpus which was defined as one million in consideration of the technologies that were available at the time as in the *Brown Corpus* and the *Survey of English Usage* Corpus. The technologies that were available then were restrictive. It is a provisional corpus and for the purpose of this study, 100 000 is enough, and the corpus will nonetheless be increased in future. However, Sinclair (1991), warns that, "This kind of corpus will be adequate for the study of the fairly frequent patterns and meanings of many thousands of words, but will not be adequate for a reliable description of the language as a whole." From the corpus compiled for this research, lexical items were selected, and their possible meanings were analysed for inclusion in the microstructure section.

Early corpus compilation projects such as the USA Brown Corpus by Brown University, Finland Helsinki Corpus and the UK Lancaster-Oslo-Bergen (LOB) and the COBUILD, for example, were not individuals' efforts. They were initiated by institutions like a university or a consortium because corpus compilation was the responsibility of institutions in these countries, considering the level of advancement of computers at that time. It was easier and quicker to compile a corpus of a million words and so the one million words became a

benchmark in early compilations. Currently, with more advanced computation, corpora run into billions like the Chinese Corpus, *zhTenTen* which had reached 15.9 billion words in November 2017 (Petrovčič, 2018). In South Africa, however, corpus building is hardly institutionalised. The University of KwaZulu-Natal and University of the Witwatersrand are the only institutions that appear to have started in this direction and have compiled corpora. The University of KwaZulu-Natal has produced over one million isiZulu tokens so far (Khumalo, 2015: 497). University of the Witwatersrand Honours students, who were the author's classmates, also compiled corpora for the following languages: isiNdebele, Setswana, isiXhosa, Siswati and isiZulu. These corpora have not yet been published and remain the property of the individual students. Still, for most official, Indigenous South African languages, corpus compilation is done by individuals, and it is not easy to access. As such, in order to develop the dictionary specified for this study and answer the first sub-question of this research which is incorporated in the overarching question - *what factors should be considered in developing a Sesotho monolingual, online, learner's dictionary* - the author compiled the corpus dubbed *Sehlalosi sa Sesotho Corpus* of 101 359 words following methods of compilation outlined in Atkins, Clear and Ostler (1992), which lessened to 13 967 types from which the +5000 words required for the dictionary were retrieved because there was no known or accessible Sesotho corpus at the initiation of this study. The current instalment of the envisaged dictionary consists of 250 lemmata and the target one will consists of +5000 lemma inclusive of the 250 all of which arise from the compiled corpus.

As already mentioned above, initially, the size of a representative and standard corpus is one million words, however, the corpus used for this research is 101 359 words and it aided this study in achieving the goal to develop a monolingual Sesotho learner's dictionary. Corpora are built with a purpose in mind, and it is the purpose that determines the size, design and nature of the corpus (Kennedy, 2014). Moreover, a corpus-based analysis is not reliant on the size of

the corpus, what is essential are the results revealed from the corpus (Leech, (1991); Krishnamurthy, (2001); McCarthy & Carter, (2001); De Schryver (2009) and Pienaar & Klerk (2009), also draw the inference that the size of a corpus does not matter, what matters is the purpose that the corpus serves. Therefore, besides the size of this corpus being smaller than the earlier standard of one million or the current large corpora, it is an encouraging contribution to Sesotho corpora, to grow the scope of the language in Corpus Linguistics.

The corpus, which has Header annotations, can be used for any linguistic, applied linguistics inquiries or to create dictionaries and the texts making it up are easily traceable due to their detailed bibliographic information presented in *TEI Lite* format. The use of a corpus as a tool in dictionary development is effective in that a corpus on its own provides words and their variants, usage examples and provides context to study the meanings of the possible lemmata.

Though the texts sampled for this corpus came from published, reputable sources, writing inconsistencies were numerous among texts. Some words in the corpus, particularly from South African written texts, do not always occur in the standard, accepted form of writing. The spelling inconsistencies interfered with the results of frequency counts from the *tlconcordancer*, one of the modules of *TLex* used for tokenizing the corpus. For instance, in one case it was found that one writer transcribed the Sesotho *ya ka* 'mine' in a different text, as *yaka* one word instead. The *tlconcordancer*, interpreted *ya ka* as two separate words and *yaka* as one word. Though both spellings are in RSA orthography, such grammar mistakes seriously affected the frequency counts. Notwithstanding, these problems have been corrected in the developed dictionary. However, differences in correctly spelled words in either orthography did not present any challenges as they were expected as in the Sesotho word for 'get out' which is *tsoa* in Lesotho orthography and *tswa* in RSA convention.

Atkins, Clear and Ostler (1992: 2-5) state that there are five stages in corpus building and these are: (i) specification and design; (ii) hardware and software; (iii) data capture and mark-up; (iv) corpus processing and (v) corpus growth and feedback. The first three stages were covered in the compilation of this corpus and as regards stage 4, only the headers were marked up using *TEI-LITE* (see 3.2.4 for details) and the body of the texts was not annotated due to the dearth of taggers for languages such as Sesotho. For the final stage in corpus building, which is growth and feedback, Atkins, Clear and Ostler (1992: 5) explain that:

> 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.

Work on stage 5 will be a continuous activity beyond this research. As matters stand, the corpus is open, hence, it is a monitor corpus. In consideration of these stages, it can be safely argued that, in the main, the developed corpus, *Sehlalosi sa Sesotho Corpu*s was built on established guidelines and so provided the data required for *Sehlalosi sa Sesotho sa Baithuti.*

The texts that were selected as detailed in 3.2.3 for this corpus are varied, they represent a number of text types and genres as they include a Sesotho prescribed workbook, exam papers, parliamentary speeches, newspaper articles, bible texts, hymns and literature. A few other texts were selected randomly based on availability and accessibility in either or both electronic and hardcopies wherein the latter were converted into electronic format. Though the goal throughout the text selection process was to create a comprehensive product suitable for learners, another objective was to end up with an electronic corpus to meet the requirement of machine-readability as one of the key characteristics of a corpus. Atkins, Clear and Ostler (1992: 3) further state that:

> Data capture is also time-consuming and the costs incurred are unavoidable and determined to a large extent by the amount of text to be captured. Printed material can be scanned by OCR devices with appropriate software. This method of conversion from print to electronic form is becoming more cost-effective each year as OCR technology improves. Alternatively, text can be keyboarded manually.

The developed corpus includes texts that were manually key-boarded to make them electronic as they were not scannable in their form. The corpus has not been appended here due to its bulky size and because the focus is the dictionary.

Using the *OCR* tool for conversion of some hardcopies and retyping those that were illegible for various reasons also forms part of corpora processing and this was possible because the computation software used, *TLex* suite, was efficient, thus ensuring that all the texts included in this corpus were machine-readable, a *sine qua non* for a corpus.

The corpus-based methodology of collecting a corpus has not only been beneficial in this study as it facilitated the making of *Sehlalosi sa Sesotho sa Baithuti* by providing the data required but also followed the desiderata for corpus compilation. Finally, using a corpus as a primary tool spared time and perhaps money, as these were cited as some of the challenges stalling the progress of Sesotho and some other South African Languages in dictionary development (Nkomo and Wababa, 2013). Otlogetswe's (2012) *Tlhalosi ya medi ya Setswana,* De Schryver (2007) *Oxford Bilingual School Dictionary: Sesotho sa Leboa-Seisemane, English-Northern Sotho* De Schryver (2010) *Oxford Bilingual School Dictionary: Zulu and English* (...) and the online *Sesotho sa Leboa (Northern Sotho) - English Dictionary* Tshwane DJe (2003) are all examples of corpus-based dictionaries of the current modern era covering some of the South Eastern Bantu languages.

5.1.2. What kind of data and tools can be used to develop a monolingual online learner's dictionary?

In this section, the discussion focusses on how the first sub-question of this research - what kind of data and tools can be used to develop a monolingual online learner's dictionary – guided the realization of the aims by focussing on the second result of this research which attempted to answer the question asked above. First, the data is under scrutiny and lastly, *TLex* Suite, as mentioned earlier includes tools that can process corpus, that is, a concordance, and a dictionary making module that can at least, accommodate the creation of monolingual and bilingual dictionaries as these are the most common types of dictionaries. For a monolingual dictionary, a lexicographer needs natural data and corpus processing tools such as a concordance and *Wordlist* generator.

A dictionary development project must have a plan that includes the type of data required and the tools that will be used to present the data. The first part of the aim of this research was to develop a monolingual dictionary for Sesotho. The three monolingual dictionaries produced in Sesotho are in print form and these are *Sehlalosi: Sesotho cultural dictionary* (1994), *Khetsi ea Sesotho* (1997), *Sethantšo sa Sesotho* (2005), and this research, additionally, aimed for this monolingual dictionary to be an online dictionary. As stated in 4.5., the prototype dictionary that has been achieved through this research is currently electronic but not yet online. The target group for this dictionary are high school learners and this is a group that is technology savvy and therefore, would be able to navigate the dictionary without much trouble.

Electronic lexicography is a viable approach to lexicography that is increasingly becoming common around the globe due to the advancement of computer technologies and software availability which is either freely available on the internet such as *Lexique Pro* which has since been replaced with *Flex* or patented software such as *TLex*. This approach can be

used to corroborate the views of scholars who have argued that lexicography is a discipline of information science (Section 2.3.1). It is the lexicographer's task to collect data, select, order, and to format the data to satisfy the user's needs. This study was guided by these principles throughout to create *Sehlalosi sa Sesotho sa Baithuti* and 5.2.1. above has discussed how the corpus-based approach facilitated data collection and selection to provide lemmata for this creation. Tools such as *TLex* make this task easier because its customizable and already has features that are relevant and required for any type of dictionary as demonstrated in (Section 4.3.2.).

As mentioned throughout this report, the data used for the dictionary is corpus data and it has proven beneficial in this research because it provided words and their different context usages, and an opportunity for the compiler to study various possible meanings through a facility named Key Word In Context (KWIC) functionality, a module of *tlCorpus* concordance. After deciding that the monolingual dictionary would be electronic, in addition to the data processing, the tools to create this dictionary were identified and studied. *TLex* Suite, which has also been used in the design of the *Oxford Punkuntšu ya Sekolo* (2007) a bi-directional North Sotho-English is a commendable tool for the dictionary designed and proved efficient for this type of dictionary (Section 3.3.2.). De Schryver and Joffe (2006: 41) state that:

> In *TshwaneLex*, a strict separation is made between the actual dictionary contents (the data), the structure of each article (the dictionary grammar or DTD (document type definition)), and the way those contents, given a certain structure, (may) look (the formatting or style). Each of those levels is fully customisable, with the data level further sub-dividable into unique and repetitive (metalanguage) material.

The ultimate goal is to put the dictionary online, and *TLex* offers upgrades that allow a compiler to move from electronic to online while also leaving the option to print the dictionary as seen with Annexure 1. Electronic resources allow for the processing, retrieval, and management of

lexicographic information and thus, the methods that were used for the existing Sesotho online dictionaries, Bukantswe *vol. 2 & 3,* could not be adopted for the nature and goals of this research (Section 2.4.) because they do not seem to have used any specific dictionary-making software. *TLex,* on the other hand, a proudly South African product is adaptable for any language, and it facilitated the creation of *Sehlalosi sa Sesotho sa Baithuti.* It has the capacity to build both monolingual and bilingual dictionaries, with the flexibility to adjust the data presentation methods as needed by a compiler. Additionally, the tools allow for further expansion of the dictionary over time, which will accommodate the growing vocabulary of a language thus the next two phases of development of this dictionary should be hassle-free.

5.1.3. Final research question

The final result arose from the third research question which reads as follows: *What kind of information can be marked up for a monolingual dictionary that can be useful for school learners of language?* Alberts (2007) explains that "Different kinds of lexicographical activities lead to different types of dictionaries. All dictionaries should, however, comply with basic structural characteristics such as a macrostructure, a microstructure and a mediostructure." All these structural characteristics of *Sehlalosi sa Sesotho sa Baithuti* were explained and presented in Section 4.3. to 4.3.2. and it was indicated how the ease of presentation of this information was aided by *TLex Suite* features. Encoding the dictionary at these three levels sorted out the presentation of the most essential information about the lemmata.

Kiango (2000) asserts that when compiling a dictionary, the compiler should identify a specific problem under each linguistic component of the language such as phonology, morphology, syntax, semantics and pragmatics etc. In this research, the phonological, morphological, and syntactic and semantic structure of Sesotho are presented in the dictionary

as demonstrated in 4.3.2. and as illustrated below in the analysis of *bajalefa*. The phonological factor of the lemmata is presented in syllabification form instead of using the IPA script for ease of comprehension by the learners. The example below contains the linguistic aspect of the lemma *bajalefa* 'heirs':

> **bajalefa** */ba-ja-le-fa/* **(=ja + lefa)** *(lebitsikgoboka)* batho ba fumanang le ho sala le mafa a kang thepa ya motho ha a hlokahala: *ke tokomane e bohlokwa hoba e tla thusa Balaodi ba Mokga ha ba lefa bajalefa ba hao ditjhelete.* {lebitso lena le bopilwe ka kutu tse pedi, kutu ya leetsi *ja* le kutu ya lebitso *lefa*.}

Morphologically, the above lemma is a compound class 2 noun as indicated by the plural class prefix *ba-* and is formed from the verb *ja* 'eat' and the noun *lefa* 'heirloom/estate' as in example *4.3.2.3.* Its singular phonological rendition is *mo-ja-le-fa* whilst semantically its gloss is 'heir'. Lastly, at the syntactic level, an example sentence is provided in this case ' *ke tokomane e bohlokwa hoba e tla thusa Balaodi ba Mokga ha ba lefa bajalefa ba hao ditjhelete*.' Further structural information is added as a *'Notes'* which explains the parts that form this compound noun.

Ali El-Sayed and Siddiek (2013: 1746) state that "Native-speaker dictionaries are for decoding activities e.g. reading and spelling". A native reader can enhance their reading and spelling by learning the sound patterns from the dictionary; they can learn to differentiate homophones by the way in which the pronunciation attribute is structured. For example, in Annexure 1, the sounds are transcribed using the syllabification method. Other Sotho language group dictionaries such as *Tlhalosi ya Medi ya Setswana* use the phonemic transcription method to mark-up pronunciation which requires the linguistic knowledge not yet available to basic education level learners. Otlogetswe (2013: 540) explains that,

> Phonemic transcriptions are important since they help users, language learners, linguists and teachers, to have a pronunciation reference point. The tonal markings in particular are important for disambiguating words which, though orthographically written the same, have different tonal distributions.

The lemmas in *Tlhalosi ya Medi ya Setswana* are word-based and not stem-based as earlier the Mabille and Mabille-Dieterlen dictionaries. *Sehlalosi sa Sesotho* has also opted for the word based approach in sorting nouns. "The word list collection used a corpus, semantic domains and the consultation of other publications such as dictionaries, grammar books and language documentations" (Otlogetswe 2013). Its alphabet stretches include digraphs and/or trigraphs-initial words. Otlogetswe (2013: 535) states that,

> *Tlhalosi ya Medi ya Setswana* was based on a 16 million token Setswana corpus (Otlogetswe 2008). The corpus is broad in its composition, with 90% of the text coming from written texts and 10% being transcribed speech. The corpus comprises novels, science texts, political texts, grammar books, poetry texts, religious texts, prose texts, spoken texts recorded from a variety of contexts such as funerals, Setswana classroom and family meetings.

Additionally, the dictionary contains both lemmas and Setswana proverbs, and in the case of proverbs there is no pronunciation mark-up. In 2013, the dictionary was consisting of 371 headwords (Otlogetswe 2013). Below is an example of the structural markers used in *Tlhalosi ya Medi ya Setswana* (2012):

> **gopê** **** /χúpε/ •*lbd*• lebadi la maina a setlhopha 15 le le tlhobosang boleng teng jwa sengwe *Ke badile pego e yotlhe mme ga go na gope fa a tlhalosang gore monna yo o dirile molato teng* (Otlogetswe, 2013)

In the above article, the lemma is bolded, and it is the same for all lemmas and proverbs in the dictionary. The asterisks indicate the frequency count from the corpus. Pronunciation mark-up is contained in forward slashes, followed by the abbreviation of the POS which precedes

definitions and in this particular definition, Otlogetswe also indicates the noun class of a word. The last item is an italicised phrase of the usage example. The lemmas in *Tlhalosi ya Medi ya Setswana* were selected based on highest frequency and further, on their semantic domains.

Another dictionary that can be compared is the *Oxford Bilingual School Dictionary: Sesotho sa Leboa-Seisemane, English-Northern Sotho* De Schryver (2007). Although it is not a monolingual dictionary, it is corpus-based and was compiled using *TLex*, moreover, it is a school dictionary. It is also word-based, with words starting from A – Y stretches, there are no letters C, Q, Z words and because of the nature of the Sesotho sa Leboa alphabet, the alphabet stretches include digraphs and trigraphs. The corpus comprised of general language texts and school textbooks. Pronunciation is marked with diacritics where applicable. Although the total of the corpus is not specified, the dictionary itself is 5000 lemmas long. Below is an illustration of the structural markers:

> **kadijela** /kadijêla/*noun 9*/- veteran. Mna. Mokgobu ke yena **kadijela** ge go etla go tša politiki. *Mr Mokgobu is a veteran when it comes to politics.* (p. 81)

In the example above, the lemma is bolded, pronunciation is marked by forward slashes and tone is indicated with diacritics. The POS and noun class are also inside slashes and written in English just before the English translation, which is followed by a usage example in Sesotho sa Leboa, and ending with the translated usage example. The lemma is further bolded within the usage example, which makes it easy for the user to identify the central word in the example. The 5000 lemmas in the *Oxford Bilingual School Dictionary: Sesotho sa Leboa-Seisemane, English-Northern Sotho* dictionary were selected based on their high frequency count.

An additional dictionary that has been used for analysis for the current study is a monolingual Sesotho dictionary by Hlalele (2005) *Sethantso sa Sesotho*, which Motjope-Mokhali et al. (2020) provides an analysis for. Hlalele (2005) did not use the corpus-based

approach but instead used an intuitive selection of lemmas. The dictionary is stem-based as shown in the example below. The alphabet comprises of letters a – z excluding letters d, g, v, which are only used in borrowed or adoptive words in *Sehlalosi* sa *Sesotho sa Baithuti* and there are also separate alphabet stretches for digraph- and trigraph-initial words. This sorting method is infamously user-unfriendly and has not been used in *Sehlalosi sa Sesotho sa Baithuti*. This sorting tradition is reminiscent of the Mabille and Mabille tradition of lexeme sorting in dictionaries and shows the impact that the forebear dictionaries have on modern lexicography. It uses Lesotho orthography as indicated in Chapter 2 and its vocabulary is loud with the absence of modern words. An example of the structural markers in provided below:

> **talim.is.a** (**.itse**) /*kutu-ketso-ketsiso*/ ho etsa hore ho talingoe … (<talima)
> (Hlalele 2005: 261)

The headword, *talima* 'look' is bolded, its past tense form is provided as shown by the suffix -*itse* in brackets, its word classification is indicated by the phrase *kutu-ketso-ketsiso* and shows that it derives from the basic word *talima* 'look' as in (<talima). A South African learner may not have problems seeing the likeliness of Lesotho *talima* with *tadima* of SA orthography, however the description of the sub-classification of this word, *kutu-ketso-ketsiso* for 'intensive verb stem', is problematic as it is not used in South Africa. This is one of the greatest cons of the Hlalele (2005) descriptions of word classifications. The POS in the slashes in Hlalele's description of an *intensive verb stem* is *kutu-ketso-ketsiso* whereas in South African Sesotho grammar, it is commonly called *kutu ya leetsi* 'verb stem'. The stem is *talim*- and the primitive verb is *talima* 'look'. Hlalele's (2005) dictionary has also added the multiple variations of the lemma.

The variation *talimisa* 'look carefully' is accommodated as well as the past intensive verbal variation. This structure is insightful, however, the one factor that is not duplicatable from Hlalele's work is the labels or titles used for POS descriptions. Since Hlalele's dictionary

was compiled based on intuitive approaches, the lemmas have no frequency indicators. *Sehlalosi sa Sesotho sa Baithuti* uses descriptions given in Doke and Mofokeng (1957) which have been largely retained in the modern era. As it is in the Setswana and Sesotho sa Leboa-English dictionaries discussed before, the lemmas in *Sethantso sa Sesotho* are bolded, as in *Sehlalosi* sa *Sesotho sa Baithuti*. This is standard practice in lexicography.

One of the major aspects that is different among the Sotho dictionaries discussed earlier and the one created in this study is the pronunciation aspect. As earlier mentioned, for Annexure 1, pronunciation was marked up using syllabification. The phonemic transcription used by Otlogetswe (2012) was discarded because learners are not familiar with the IPA system because it is not part of their school curriculum. And Sesotho sa Leboa uses circumflexes whereas these are not used in RSA Sesotho writing and so they were deliberately avoided as they could cause confusion. Since the learners are functionally literate by the time, they start using a dictionary, the definitions and usage examples can help them to determine which meaning is applicable and what the correct intonation should be. Thus, syllabification was used instead of phonemic transcriptions or circumflexes in *Sehlalosi sa Sesotho sa Baithuti*.

Bartlett, Kondrak and Cherry (2009, p. 308) state that "Syllabification is the process of dividing a word into its constituent syllables. [...] syllables are, technically speaking, phonological entities that can only be composed of strings of phonemes." In a language such as Sesotho, which has seven phonemic vowels which are /*a, e, e, i, o, o, u*/. The mid-vowels /e/ and /o/, are four in Sesotho because of the different heights and whether the articulation point is at the back or the front of the oral cavity, and each is represented simply as /e/ or /o/ in writing. Though this is a source of difficulty in decoding the meaning of any word or homonym in which they occur, it is particularly important that they be marked up by either using a diacritic such as circumflex or IPA for disambiguation purposes, however, this is not done here for the reason stated in the preceding paragraph above practicality, simplicity and knowledge

level of users. In instances such as the ones in the example below, phonemic transcriptions would be helpful:

Example 5.2.3.1:

bona ‘see’ and *bona* ‘them’
bobe ‘ugliness’ and *bobe* ‘anticipate’

With *bona* ‘see’ the [o] is a /ɔ/ in IPA and in *bona* ‘them’ it is a [o]. The [e] in *bobe* ‘ugliness’ is /i/ in IPA and is /ɛ/ in *bobe* ‘anticipate’. The phonemic transcriptions would make it easier to disambiguate nonetheless, the addition of usage example is one method by which learners would be able to tell which sense of the word is being used according to context and therefore, determine the tonation.

In addition to reading and spelling, from the dictionary entries in Annexure 1, one can also learn grammar for example, taking an article such as *aba* in Figure 4.3.2.8., which is a primitive verb with multiple derivations. The primitive has a usage example and so do all its inserted derivation, a learner can deduce how the change of tense, or the extensions of the verb influence the structure of the sentences where its used.

Hanks (1979) asserts that a lexicographer must be proactive when creating dictionary entries i.e., the entries should contain all the information that a user would inquire about when consulting a dictionary. Monolingual dictionaries are primarily meant for native speakers of a language, as a result, such a dictionary must reveal the structure and use of a language, by presenting the linguistic aspects of each lemma (cf. Section 5.2.1.). This can be seen even with one of the monolingual online dictionaries that was created with *TLex*, the *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete* (2004), a North Sotho general monolingual dictionary. The creators of *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete* have not added any sort of pronunciation or sound mark-up for certain words but have consistently annotated lemmas with

POS, noun class, definitions, usage examples and cross-references and notes were applicable (https://africanlanguages.com/psl/).

Even though *Tlhalosi ya Medi ya Sesotho (2012)* and *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete (2004)* are not learner's dictionaries, they are monolingual and comparing them to the dictionary compiled herein, the basic information to add in a monolingual dictionary seem to be morphological, syntactic, and semantic information. Varantola (2002) cautions that we should also be wary of how meanings are presented in a dictionary, because this will determine user-friendliness of the dictionary and its usefulness largely.

Finally, a closer look at *Sehlalosi sa Sesotho sa Baithuti* in comparison with the online dictionary, *Bukantswe v.3.* Below is a table juxtaposing *Sehlalosi sa Sesotho sa Baithuti* with *Bukantswe v.3,* which is an online Sesotho-English dictionary and the only Sesotho online dictionary that has reference information.

Bukantswe v.3* vs. *Sehlalosi sa Sesotho sa Baithuti

ANALYSIS	**BUKANTSWE V.3 Olivier, 2009**	**SEHLALOSI SA SESOTHO SA BAITHUTI MJ Mahloane, 2022 (this research)**
Lemma selection	Unknown	Corpus-driven. Selected from a corpus of 101 359
Lemma treatment	Word-based but search results include phrases Bilingual Eng-Sesotho	Nouns are word-based and verbs, qualificatives and adverbs are stem-based. There are both primitive and derived, as well as loan words. Monolingual
Alphabet stretches	A – Z	A – B monographs (First transaction on 250 lemmas)
Alphabet arrangement	Queries in either language	A – B
Use of corpora	Grammatical number, sg-pl, always inconsistently indicated Unclear what the source of the lexemes is.	Grammatical number, singular and plural is indicated. The corpus was used for extraction of lemmas; usage examples and for semantic analysis.
Information	Homonyms indicated	Homonyms also indicated
Columnar Structure	One-column structure	One-column structure
Language(s)	Sesotho-English and lemmas in SA Sesotho	Sesotho only and lemmas in South African orthography
Pronunciation	Not marked in any form	Lexemes syllabified for pronunciation purposes included
No. of Lemmata	10075	250
No. of Lexemes per page	N/A	N/A
Structural markers	Sesotho English /nka tsamaya// I can walk	bajalefa /ba-ja-le-fa/ (=ja + lefa) (lebitsokgoboka) batho ba fumanang le ho sala le mafa a kang thepa ya motho ha a hlokahala: ke tokomane e bohlokwa hoba e tla thusa Balaodi ba Mokga ha ba lefa bajalefa ba hao ditjhelete. {lebitso lena le bopilwe ka kutu tse pedi, kutu ya leetsi ja 'eat' le kutu ya lebitso lefa 'estate'}
Frequency indicators	Unknown	High frequency count in the case of concords and middle and low frequency selection in the case of POSs

Table 5.1.3.1: *Comparison of Headwords/Lemmata (Macrostructure) with online dictionaries*

5.2 Results: Significance, practicality of study and recommendations.

The function of any dictionary is to serve as a reference resource. Dictionaries convey the status of words and meanings and usages. Hanks (2010: 581) observes that "words have the potential to make meanings when put into context – and they are associated with particular sets of syntagmatic patterns, which can be discussed through painstaking corpus analysis." Hanks

(2010) identifies that there are two major kinds of monolingual dictionaries; these are the tradition-based and the synchronic dictionary. The tradition-based dictionary contains the history and etymology, i.e., it indicates how the word has evolved over time to what it has come to mean in the present day – it is diachronic. On the other hand, "the focus in a synchronic dictionary is on reporting conventional meanings and use, rather than on historical and etymological research" (Hanks, 2010: 582).

A prototype dictionary dubbed *Sehlalosi sa Baithuti sa Sesotho* has been developed as the result of this research and presented here, it has both synchronic and diachronic attributes. Etymological information is provided for lemmas that required it because *TLex* makes provision for the addition of etymological information for each lemma in the article structure. It is also synchronic in that it provides lemma definitions and usage examples from various types of texts from varying periods of writing.

As stated before, when this research was initiated, there was no accessible Sesotho corpus readily available and therefore, a corpus of 101 359 words was created by the author, which comprises of two Sesotho orthographies. The corpus is annotated at header level, and the variety of text samples it covers have been organized according to genre, thus subject specific glossaries could be developed from it making the corpus more practicable than it could have been had this approach not been used.

Concords, also known as agreement markers - which establish agreement between the subject or object noun with the pronoun, the qualificative and verbs - and tense markers have the highest frequency than the lexical and fixed categories thus confirming their centrality and importance in Sesotho morphology and that of its sister languages - Setswana, North Sotho and Nguni - and the Bantu family at large. Though concords are not parts of speech by themselves,

they are an important part of any part of speech they prefix to and so need to be included in any dictionary.

It is recommended that Sesotho scholars should look into the development of a larger corpus and focus particularly on its annotation. Tools such as *TEI-lite,* which are freeware are available to simplify the annotation process. Having an annotated corpus for Sesotho will further enable the development of Human Language Technologies for lesser-known languages such as Sesotho.

Additionally, the PanSALB should intensify the dissemination of standardised spelling and orthography rules for Sesotho completed by its National Language Body among the speech community across the country to avoid further inconsistencies in future writings and in speech alike.

5.3. Implications of Study

For this research, the corpus that was compiled as well as the prototype dictionary created alongside it, are a contribution not only to Sesotho lexicography and Sesotho corpus studies but also to other lesser-studied languages in South Africa and Africa in general. It will also serve as a template for a larger Sesotho monolingual online learner's dictionary. Moreover, the study is adding to the field of metalexicography as it is based on insights from other researchers as they apply to Sesotho. The result of this research is a +250 strong lemmata prototype Sesotho monolingual learner's dictionary, which will be used as a template to create the idealised c.a. 5000 lemmas dictionary project. Through the process of creating the prototype dictionary, the different language processing tools were identified. Tools such *TshwaneLex* which encompass corpus compilation features, concordances and wordlist generators that are adaptable for South African languages are brought into the limelight. Further, these tools have lessened

significantly the amount of time spent on compilation of the lexicon and the actual compilation of the dictionary itself. The monitor corpus compiled is also a resource that shall be reused for other research purposes and shall be expanded.

There is no evidence that Sesotho lexicographers used corpus approaches for the already existing dictionaries as corpus studies are an off shoot of disappointing machine translation efforts of the 1960's and later. Dictionary-making by using compiled corpora is now firmly established as a lexicon compilation method as opposed to earlier methods such as using cardexes and shoe boxes. Also, corpora have a wider reach of the lexicon of a language as compared to all other previously used methods. The bigger and varied the corpus, the better the language coverage. Therefore, the use of Corpus Linguistics in this study introduces alternative methods that Sesotho dictionary makers could explore for future work. A machine-readable corpus, as a basic tool in Corpus Linguistics, once compiled, has proved to be convertible to the needs of language research and practice. For example, in this study, a corpus became a source from which the use of words and the inconsistencies in the structure of some words were observed. From these observations, it is recognised that in addition to Sesotho requiring lexicographic materials, other language resources such as standard grammar books have to be developed and popularised among the speakers.

Lastly, as shown in both chapters 1 and 2, Sesotho has a rich tradition of bilingual lexicography spanning over 160 years and most of it originated in Lesotho. However, not only is its monolingual dictionaries effort sparse but it also needs to keep pace with the development of Sesotho language. It is anticipated that this prototype is a step in that direction and will have a great influence on the Basotho in both the Republic of South Africa and Lesotho to step up lexicographic efforts to enhance the current ones and lessen the spelling confusions resulting from the two parallel orthographies of Sesotho.

5.4 Limitations

The first major limitation was the paucity of Sesotho corpora, the preferred method for this research. Notwithstanding, the author decided to develop corpora for the dictionary project and as corpora development takes a lot of time, a significant amount of time was spent on this aspect and affected the size of the corpus compiled for this research. Though +100 000 may look small, it is a feat considering that it was a one-man effort and there were very few Sesotho electronic texts available at the initiation of this research which meant a lot of scanning and converting texts to acceptable format for the analysis tools. Secondly, for a start-up dictionary like this one, a +100 000 word corpus is thought to be good for lesser-known languages.

The second limitation is the paucity and unavailability of Sesotho metalexicography and unavailability/inaccessibility of most dictionaries listed in chapter 2. There is plenty of theory on lexicography for the world's major languages such as English and French but for lesser-known languages such as Sesotho, theory is only budding in the latest part of the 20th century. A researcher needs to study available dictionaries to discern theoretical issues relevant to Sesotho. Most of the published Sesotho dictionaries followed the Mabille and Mabille and Dieterlen tradition of dictionary-making without trying to enrich it. Also, most of the surveyed dictionaries were not even available in the *OPAC* system of libraries and so the researcher's effort was limited in that sense or depended on secondary sources only.

Although a prototype dictionary has been created as was the goal of this research, it has been shown that the corpus used to harvest lemmas does not cater for the creation of monolingual dictionaries to full measure. For instance, the distinguish feature of a monolingual dictionary is that it provides definitions of words. However, one cannot find explicit definitions of words from a corpus, rather, the lexicographer can study the contexts of words and form their meaning. Thus, involving actual speakers or linguistics in a dictionary-making project

could enhance the weight and applicability of the definitions assigned, and in this research, other linguistics or lexicographers were not involved due to time restrictions lost to compiling the corpus, the bloodline of this prototype, but these aspects will be considered in the next segment of the dictionary project.

Nonetheless, for defence of this study, the dictionary was a pilot project to indicate that there is a need for other types of dictionaries in Sesotho and that there are methodologies available to assist in expanding the scope of Sesotho lexicography.

5.5 Future Work

Future work on this study will focus on populating the microstructure of the rest of the ±4750 lemmas with required information to have the initially desired monolingual learner's dictionary, and this shall be followed by the upgrading of the software version to one which allows permission for publication and once that is done, a *URL* will be created, and the dictionary will be put online.

Scholars have also identified that many times, lexicographers use words that are not listed in the dictionary when explaining the listed lemmas. As a result, further work will look into adding words that are used in the definition of current lemmas to the lexical list and define those as well.

5.6. Conclusion of study

This final section sums up the study under the indicated subtitles.

5.6.1. Brief statement of the problem

At the start of this study, one could not find following types of dictionaries in Sesotho: learner's dictionaries, monolingual print or online dictionaries, or monolingual learner's dictionaries. It

is for this reason that this research initiated the investigation reported herein to close this gap. According to De Kind et al. (2012:116), lexicography was introduced to African Languages about 360 years ago with the *Vocabularium, Latinum, Hispanicum, e Congense* in 1652 by Van Gheel (De Kind et al., 2012) as the precursor. Sesotho lexicography is over 160 years old and most of its dictionaries except for *Sehlalosi* (Matšela, 1991), *Sethantso sa Sesotho* (Hlalele, 2005), and *Khetsi ea Sesotho* (Pitso, 1997) are either bilingual or multilingual. Section 1.2. listed the known Sesotho Dictionaries. The first comprehensive Sesotho dictionary written in RSA spelling – *Southern Sotho-English Dictionary* - was produced in 1974 by Morija Publishers and subsequently the lexicographic tradition in South Africa remained bilingual. A survey of Sesotho dictionaries has revealed that not only are there no monolingual learner's dictionaries, but also, there are no monolingual learner's online dictionaries.

5.6.2. Aim and Research Questions

The aim of this study was to develop a monolingual online Sesotho Learner's Dictionary using Corpus Linguistics approaches, specifically, the corpus-based approach. By setting on to develop this dictionary, the following question, which helped pave the way for the methods that would be used to develop the dictionary and the analysis thereof was: *What are the factors of consideration when developing a dictionary with these specifications?* This question was further broken down into the following: *What kind of data and tools can be used to develop a monolingual online learner's dictionary? What kind of information can be marked up for a monolingual dictionary that can be useful for language learning?*

5.6.3. Rationale, Goal(s)

This research was driven by the understanding that a dictionary is a basic resource to learning and a monolingual one is mostly critical in the learning of Sesotho as a subject for L1 speakers. Language learners require dictionaries. Additionally, the world at large has moved to advanced

human language technologies and so while the paper dictionary is useful and will remain so for other users, an online dictionary is unavoidable in keeping with the recent and future times. Further, this research studied other Sesotho dictionaries and the corpus itself and identified serious spelling inconsistencies across and so the availability of such a resource will promote the standardised writing conventions of South African Sesotho while not precluding Lesotho orthography as the differences in writing between the two is minimal. The dictionary sought to use the current standard Sesotho spelling as advocated by PanSALB to address the inconsistencies discovered.

The ultimate goal is to produce a monolingual online Sesotho Learner's Dictionary of +5000 lemmas, a realistic size for learners, to be used primarily by school learners as a reference resource. However, a dictionary of that size is too big a project for a Master's research and so in order to achieve this goal, the intended work was divided into three phases – development of the corpus and the prototype, populating the macrostructure of the +4750 remaining lemmas of the dictionary later and finally customizing it to an online resource. As a result, the immediate goal is to process the first phase, developing a corpus and prototype that will serve as a model of the entire effort.

5.6.4. Methods used for analysis

The methods used in this research are both quantitative and qualitative, which are both accommodated in the corpus-based approach. The quantitative part involved sampling texts for the corpus, generating a wordlist and studying word frequency counts, and selecting lemmas for the dictionary – it was applied on the macrostructure of the dictionary. The qualitative methods on the other hand, applied to the microstructure of the dictionary where the structure of lemmas was studied and aspects such as appropriability and spelling were checked; the contexts of the lemmas was studied from the corpus to see the number of senses or POS they

occupied in order to assign them the relevant definitions, and for extraction of usage examples. Further, the compiler is a 1st Language Speaker of Sesotho, and the knowledge of the language was also applied when assigning examples and definitions, and assessing the spelling of the words etc. For the dictionary microstructure, the semasiological approach to lexicography was applied.

REFERENCE LIST

Alberts, M., 2001. Lexicography versus terminography. *Lexikos*, *11*.

Alberts, M., 2007. Dictionary structures and dictionary using skills. *PanSALB, Afrilex*.

Alberts, M. and Mollema, N., 2013. Developing legal terminology in African languages as aid to the court interpreter: A South African perspective. *Lexikos*, *23*, pp.29-58

El-Sayed, A.N.A.A.A. and Siddiek, A.G., 2013. Monolingual & Bilingual Dictionaries as Effective Tools of the Management of English Language Education. *Theory & Practice in Language Studies*, *3*(10).

Anthony, L., 2014. AntConc (Version 3.4. 3)[Computer Software]. Tokyo, Japan: Waseda University.

Assam, B.N., 2017. Planning a dictionary for mother tongue education: a conceptual framework for Gabonese languages. *Lexikos*, *27*, pp.378-402.

Atkins, S., Clear, J. and Ostler, N., 1992. Corpus design criteria. *Literary and linguistic computing*, *7*(1), pp.1-16.

Atkins, B.S. and Rundell, M., 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Baldinger, K., 1980. Semantic theory: towards a modern semantics.

Bartlett, S., Kondrak, G. and Cherry, C., 2009, June. On the syllabification of phonemes. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 308-316).

Bennett, J., Tsoeu, N. and Buthelezi, T., 2006. *Multilingual Illustrated Dictionary: English, IsiZulu, Sesotho, IsiXhosa, Setswana, Afrikaans, Sepedi*. Pharos.

Bergenholtz, H. and Kaufmann, U., 1997. Terminography and lexicography. A critical survey of dictionaries from a single specialised field. *HERMES-Journal of Language and Communication in Business*, (18), pp.91-125.

Bergenholtz, H., Cantell, I., Fjeld, R.V., Gundersen, D., Jónsson, J.H., and Svensén, B., 1997. *NLO. Nordisk leksikografisk ordbok*. Oslo: Universitetsforlaget.

Bergenholtz, H. and Gouws, R., 2010. A new perspective on the access process. *HERMES-Journal of Language and Communication in Business*, (44), pp.103-127.

Bergenholtz, H. and Gouws, R.H., 2012. What is lexicography? *Lexikos*, *22*, pp.31-42.

Bergenholtz, H. and Gouws, R.H., 2013. The Presentation of Word Formation in General Monolingual Dictionaries. *Lexikos*, *23*, pp.59-76.

Biber, D., 1993. Representativeness in corpus design. *Literary and linguistic computing*, *8*(4), pp.243-257.

Biber, D., Douglas, B., Conrad, S. and Reppen, R., 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., 2010. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford handbook of linguistic analysis*.

Boas, F., 1940. Race, Language, and Culture. Chicago, London: The University of Chicago Press.

Bowern, C. and Dwyer, A. 2012. Lexicography 4. Colang: Institute for Collaborative Research, pp. 19-22.

Burkhanov, I.Y., 1998. *Lexicography: A dictionary of basic terminology*. Wydawn. Wyższej Szkoły Pedagogicznej w Rzeszowie.

Bumard, L., 2000. Reference Guide for the British National Corpus: World Edition. *Retrieved Febmary*, *16*, p.2009.

Cabré, M.T., 1999. *Terminology: Theory, methods, and applications* (Vol. 1). John Benjamins Publishing.

Census, 2011.

Čechová, M.Z., Microstructure of Learner's Dictionaries.

Chomsky, N., 2000. *New horizons in the study of language and mind*. Cambridge University Press.

Considine, J. ed., 2019. *The Cambridge World History of Lexicography*. Cambridge University Press.

Corbett, G.G., 1991. *Gender.* Cambridge University Press.

Davies, A. and Elder, C., 2004. *The handbook of applied linguistics*. By Blackwell Publishing Ltd.

De Kind, J., De Schryver, G.M. and Bostoen, K., 2012. Pushing back the origin of Bantu lexicography: The Vocabularium Congense of 1652, 1928, 2012. *Lexikos*, *22*, p.159.

Demuth, K., 2000. Bantu noun class systems: Loan word and acquisition evidence of semantic productivity. Systems of nominal classification, pp.270-292.

Demuth, K., 2003. The acquisition of Bantu languages. *The Bantu languages*, *209*, p.222.

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*, *18*(1-4), pp.89-106.

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

De Schryver, G.M. and Joffe, D., 2006. The users and uses of TshwaneLex One. In *4th International Workshop on Dictionary Writing Systems (DWS-2006)* (pp. 41-46). (SF) 2 Press.

De Schryver, G.M., 2007. Oxford bilingual school dictionary: Northern Sotho and English.

De Schryver, G.M. and De Pauw, G., 2007. Dictionary Writing System (DWS)+ Corpus Query Package (CQP): The Case of" TshwaneLex". *Lexikos*, *17*.

De Schryver, G.M. and E. Taljard., 2007. Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho. Lexikos 17: 37-5.

De Schryver, G. and Taljard, E., 2007. Compiling a corpus-based dictionary grammar: An example for Northern Sotho. *Lexikos*, *17*.

de Schryver, G., 2009. Lexikos at eighteen: An analysis. *Lexikos*, *19*.

Doke, C.M. and Mofokeng, S.M., 1957. *Textbook of southern Sotho grammar*. Longmans, Green.

Doke, C.M. and Mofokeng, S.M., 1974. *Textbook of Southern Sotho grammar*. Johannesburg: Longmans. xv+491pp.

Durkin, P. ed., 2016. *The Oxford handbook of lexicography*. Oxford University Press.

Casalis, E., 1841. *Études sur la langue sechuana*. Imprimerie royale.

Fillmore, C.J., Kay, P. and O'connor, M.C., 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pp.501-538.

Furiassi, C., 2003. False Anglicisms in Italian Monolingual Dictionaries: A Case study of some electronic editions. *International Journal of Lexicography*, *16*(2), pp.121-142.

Fort, K., 2011. Corpus Linguistics: corpora.

Gast, V., 2006. The distribution of also and too: a preliminary corpus study. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), pp.163-176.

Gouws, R.H. 1996. A Sequence for Meeting Lexicographic Needs. *Lexicography as a Financial Asset in Multilingual South Africa*: 97-110. Pretoria: Department of Arts, Culture, Science and Technology.

Gouws, R.H. and Prinsloo, D.J., 2005. Left-expanded article structures in Bantu with special reference to IsiZulu and Sepedi. International journal of lexicography, 18(1), pp.25-46.

Gouws, R.H., 2007, December. A critical look at the presentation of. In *Dictionary Visions, Research and Practice: Selected papers from the 12th International Symposium on Lexicography, Copenhagen 2004* (Vol. 10, p. 55). John Benjamins Publishing.

Guillot, M.N. and Kenning, M.M., 1994. Electronic monolingual dictionaries as language learning aids: a case study. *Computers & Education*, *23*(1-2), pp.63-74.

Hair, P.E.H., 1970. The contribution of early linguistic material to the history of West Africa. *Language and History in Africa*, pp.50-63.

Hamel, L., 1965. English-Southern Sesotho Dictionary (Vol. 1). Catholic Centre.

Hanks, P., 1979. To what extent does a dictionary definition define? *ITL-International Journal of Applied Linguistics*, *45*(1), pp.32-38.

Hanks, P., 2010. Compiling a monolingual dictionary for native speakers. *Lexikos*, *20*, pp.580-598.

Hanks, P., 2012. The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), pp.398-436.

Hartmann, R.R.K. ed., 1983. *Lexicography. Principles and practice (Applied Linguistics)*. London/New York: Academic Press, p.13.

Hartmann, R.R.K. ed., 1990. *Lexicography in Africa*. Exeter: University of Exeter Press.

Hartmann, R.R.K., 1996. Lexical culture. *Lexikos*, *6*, pp.184-198.

Hartmann, R.R.K., 2001. *Teaching and Researching Lexicography*. Harlow, Longman.

Hartmann, R.R.K., 2008, July. Twenty-five years of dictionary research: Taking stock of conferences and other lexicographic events since LEXeter'83. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008) ed. by Elisenda Bernal & Janet DeCesaris. Barcelona* (pp. 131-148).

Heid, U. and Gouws, R.H., 2006. A model for a multifunctional dictionary of collocations. In *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-9 settembre 2006* (pp. 979-988).

Heid, U., 2014. Natural Language Processing techniques for improved user-friendliness of electronic dictionaries. In *of the XVI Euralex International Congress: The User in Focus* (pp. 47-62).

Herbst, T., 1996. What are collocations: sandy beaches or false teeth?

Hiswelókë's Sindarin Dictionary (HSD) Project (2007)

Hlalele, B., 2005. *Sethantso sa Sesotho*. Maseru: Longman.

Howard, J., 2002. Lexicography: An Introduction. *L&NY: Taylor&Francis Routledge*.

https://sketchengine.eu/-chinese-corpus/

http://www.laurenceanthony.net/software/antconc/

http://lexically.net/

http://www.lexiquepro.com/index.htm

https://tshwanedje.com/corpus/

https://tshwanedje.com/tshwanelex/

https://www.gov.ls/document-category/official-speeches/

https://africanlanguages.com/psl/

Hyland, K., Huat, C.M. and Handford, M. eds., 2012. *Corpus applications in applied linguistics*. A&C Black.

Imbayarwo, T., 2008. A Biblical Hebrew lexicon for translators based on recent developments in theoretical lexicography.

Jackson, H., 2002. *Lexicography: an introduction*. Psychology Press.

Joffe, D., De Schryver, G.M. and Prinsloo, D.J., 2003. Computational features of the dictionary application" TshwaneLex". *Southern African linguistics and applied language studies*, *21*(4), pp.239-250.

Kaplan, R.B. and W. Grabe, 2002. A modern history of written discourse analysis. *Journal of Second Language Writing*, 11, Issue 3, pp.191-223.

Katz, J.J. and Fodor, J.A., 1963. The structure of a semantic theory. *language*, *39*(2), pp.170-210.

Kawalya, D. and De Schryver, G.M., 2013. Introducing a new lexicographical model: alphaconceptual+ (and how it could be applied to dictionaries for Luganda). *Lexikos*, *23*, pp.172-200.

Kennedy, G., 1998. *An Introduction to Corpus Linguistics (Studies in Language and Linguistics).* London: Addison Wesley Longman.

Kennedy, G., 2014. *An introduction to corpus linguistics*. Routledge.

Khairuddin, S. Abdurabuh, E. Diana, N. Rabi, M. Noor, N. 2011. Corpus Linguistics in Lexicography. Available at: http://www.slideshare.net/ayfa/sinopsis. Accessed 26 April 2014.

Khumalo, L., 2015. Semi-automatic term extraction for an isiZulu linguistic terms dictionary. *Lexikos*, *25*, pp.495-506.

Kiango, J.G., 2000. *Bantu lexicography: a critical survey of the principles and process of constructing dictionary entries*. Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Kosch, I., 2012. Challenges of predictability and consistency in the first comprehensive Sotho dictionary. *Lexikos*, *22*, pp.226-242.

Kosch, I., 2013. An Analysis of the Oxford Bilingual School Dictionary: Northern Sotho and English (De Schryver 2007). *Lexikos*, *23*, pp.611-627.

Kriel, T.J., 1958. *The New Sesotho-English Dictionary*. Afrikaanse Pers.

Krishnamurthy, R., 2001. Size matters: Creating dictionaries from the world's largest corpus. In *KOTESOL Proceedings 2000* (p. 169).

Krishnamurthy, R., 2006. Corpus lexicography. In Encyclopedia of language & linguistics (pp. 250-254). Elsevier.

Krishnamurthy, R., 1987. *1987-The Cobuild Dictionary*. [online] Researchgate. Available at: <https://www.researchgate.net/publication/265681918_1987 The_Cobuild_Dictionary_1987/citation/download> [Accessed 23 February 2014].

Kytö, M. and Rissanen, M., 1993. “By and by enters [this] my artificiall foole... who, when Jack beheld, sodainely he flew at him”: Searching for syntactic constructions in the Helsinki. *Early English in the computer age: explorations through the Helsinki Corpus*, (11), p.253.

Leech, G.N., 1991. *The state of the art in corpus linguistics*. na.

Leech, G.N., 1992. Corpora and theories of linguistic performance. *Directions in corpus linguistics, 1992*, pp.105-122.

Leech, G.N., 2011. Corpora and theories of linguistic performance. In *Directions in corpus linguistics* (pp. 105-126). De Gruyter Mouton.

Lehmann, W.P., 2013. *Historical linguistics: an introduction*. Routledge.

Lehmann, C., n.d. *Lexicography*. [online] Christianlehmann.eu. Available at: <https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html?https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/microstructure.html> [Accessed 3 June 2020].

Louwrens, L.J., 1997. On the development of scientific terminology in African languages: the terminographer's dilemma in a new dispensation. *Lexikos*, *7*, pp.245-251.

Mabille, A. and Dieterlen, H., 1974. Southern Sotho-English Dictionary Revised by RA Paroz. *Morija, Morija Sesuto Book Depot*.

Mabille, A. and Dieterlen, H., 1979. *Sesuto-English dictionary*. 9th ed. Morija: Morija Museum and Archives.

Mangeot, M. and Sérasset, G., 2001, November. Papillon lexical database project: Monolingual dictionaries and interlingual links. In *NLPRS-2001* (pp. 119-125).

Mavoungou, P.A., 2010. *Lexicographie et confection des dictionnaires au Gabon*. AFRICAN SUN MeDIA.

McArthur, T. and MacArthur, T., 1998. *Living words: language, lexicography, and the knowledge revolution*. University of Exeter Press.

McCarthy, M. and Carter, R., 2001. Size isn't everything: Spoken English, corpus, and the classroom. *Tesol Quarterly*, *35*(2), pp.337-340.

McEnery, A.M., 1992. *Computational Linguistics*, Wilmslow: Sigma Press, p.25.

McEnery, T.W. and Wilson, A., 2001. A.(1996) Corpus Linguistics.

McEnery, T.W. and Wilson, A., 2001. *Corpus Linguistics: An Introduction*. Edinburg: Edinburg University Press, pp.1-4.

Mchombo, S., 1998. Chichewa. In Spencer, Andrew & Arnold Zwicky (eds.), The Handbook of Morphology, pp.500-520. Oxford: Blackwell.

Mchombo, S.A., 2017. Chichewa (Bantu). *The handbook of morphology*, pp.500-520.

Macmillandictionary.com. 2021. *E-DICTIONARY (noun) definition and synonyms | Macmillan Dictionary*. [online] Available at: <https://www.macmillandictionary.com/dictionary/british/e-dictionary> [Accessed 8 September 2017].

Masiea, R.J.R., 1981. *Seyalemothati*. Van Schaik Publishers.

Matlosa, S., 1982. *Mopheme*. Mazenod Book Centre.

Mdee, J.S., 1997. Language learner's use of a bilingual dictionary: a comparative study of dictionary use and needs. *Lexikos*, *7*, pp.94-106.

Meinhof, C. 1932. *Introduction to the phonology of the Bantu languages*. Berlin: Dietrich Reimer/Ernst Vohsen.

Moeketsi, R.H., 1991. *The Concise Trilingual Pocket Dictionary: English, Southern Sotho, Afrikaans*. Ad Donker Publisher: Johannesburg.

Mongwe, M.J., 2006. *The role of the South African national lexicography units in the planning and compilation of multifunctional bilingual dictionaries* (Doctoral dissertation, Stellenbosch: University of Stellenbosch).

Mojela, V.M., 2013. A balanced and representative corpus: the effects of+ strict corpus-based dictionary compilation in Sesotho sa Leboa. *Lexikos*, *23*, pp.286-296.

Moleleki, M.A., 1999. The state of lexicography in Sesotho. *Lexikos*, *9*(9), pp.241-247.

Motjope-Mokhali, T., Kosch, I.M. and Mafela, M.J., 2020. Dictionary usage in teaching a mother tongue with reference to Sesotho. *South African Journal of African Languages*, *40*(3), pp.335-341.

Motsapi, M., 2015. *English-Sesotho Dictionary*. 1st ed. Bedfordview: South African Heritage Publishers.

Neubauer, F., 1984. The language of explanation in monolingual dictionaries. *LEX'eter'83 Proceedings. Tiibingen: Max Niemeyer Verlag*, pp.117-123.

Neubauer, F., 2017, December. The language of explanation in monolingual dictionaries. In *LEXeter'83: proceedings* (pp. 117-123). Max Niemeyer Verlag.

Nesi, H. and Durkin, P., 2015. The demands of users and the publishing world: printed or online, free or paid for?. *The oxford handbook of lexicography*, pp.579-589.

Nkomo, D. and Wababa, Z., 2013. IsiXhosa lexicography: past, present and future. *Lexikos*, *23*, pp.348-370.

Olivier, J., 2006. *Sesotho dictionary - Bukantswe Online v.2*. [online] Sesotho.org. Available at: <http://www.sesotho.org/dictionary/> [Accessed 10 April 2014].

Olivier, J., 2009. *Bukantswe 3.1 - Sesotho.org*. [online] Bukantswe.sesotho.org. Available at: <http://bukantswe.sesotho.org/> [Accessed 10 April 2014].

Otlogetswe, T.J., 2013. Introducing Tlhalosi ya Medi ya Setswana: The design and compilation of a monolingual Setswana dictionary. *Lexikos*, *23*, pp.532-547.

Patton, M.Q. and Cochran, M., 2002. A guide to using qualitative research methodology.

Petrovčič, M., 2017. zhTenTen – Chinese corpus from the web | Sketch Engine. [online] Sketch Engine. Available at: <https://www.sketchengine.eu/zhtenten-chinese-corpus/> [Accessed 16 June 2018].

Pienaar, L. and De Klerk, V., 2009. Towards a corpus of South African English: corralling the sub-varieties. *Lexikos, 19*.

Pitso, T.T.E., 1997. *Khetsi ea Sesotho*. Cape Town: CTP Book Printers.

Podhajecka, M. 2004. Historical Development of Lexicography Genres: Some Methodological Issues. Somerville, Cascadilla Proceedings Project.

Prinsloo, D.J. and Heid, U. 2011. A bilingual dictionary for specific user group: Supporting Setswana speakers in the production and reception of English, South African Journal of African Languages, 31 (1), pp. 66-86.

Prinsloo, D.J. and de Schryver, G. 2001. Corpus applications for the African languages, with special reference to research, teaching, learning and software. Southern African Linguistics and Applied Language Studies, 19:1-2, pp. 111-131.

Prinsloo, D.J. and Mashamaite, K.J., 1999. Sepedi dictionary project. *Lexikos*, *9*(9), pp.231-233.

Rissanen, M., Kyto, M., Kytö, M. and Palander-Collin, M. eds., 1993. Early English in the computer age: Explorations through the Helsinki Corpus (No. 11). Walter de Gruyter.

Rey, A. 1995. Essays on terminology. Amsterdam: John Benjamins Publishing Company.

Sager, J. 1990. A practical course in terminology processing. Amsterdam: John Benjamins Publishing Company.

Sampson, G. and McCarthy, D. eds., 2005. *Corpus linguistics: Readings in a widening discipline*. A&C Black.

Scott, M., 2010. WordSmith tools help. *Liverpool: Lexical Analysis Software*.

Setaka, M. and Prinsloo, D.J., 2020. A Critical Evaluation of Three Sesotho Dictionaries. *Lexikos*, *30*, pp.445-469.

Sinclair, J.M., 1985. Lexicographic evidence. *Dictionaries, Lexicography and Language Learning. Pergamon Press, New York, NY*.

Sinclair, J.M., 1987. The dictionary of the future. *Library Review*.

Sinclair, J.M. and Sinclair, L., 1991. *Corpus, concordance, collocation*. Oxford University Press, USA.

Singh, R.A., 1982. *An introduction to lexicography* (No. 26). Central Institute of Indian Languages.

Smit, M., 2000. Using frames in special-field lexicography: An ethnomusicological case study. *Lexikos*, *10*(10), pp.173-188.

Spencer, A., 1998. Bulgarian clitics-the limits of Optimality.

Tegnidisc, 1966. Southern Sotho dictionary = Suid-Sotho woordeboek. Stellenboch.

Thetso, M.L., 2018. *The Distribution and Interpretation of the Qualificative in SeSotho* (Doctoral dissertation).

Tognini-Bonelli, E., 2001. *Corpus linguistics at work* (Vol. 6). John Benjamins Publishing.

Turner, Chris. (2015). Re: What is the difference between Word Type and Token? Retrieved from:https://www.researchgate.net/post/What_is_the_difference_between_Word_Type_and_Token/5640fb975dbbbd13038b4586/citation/download.

Van Dijk, T.A., 1985. Handbook of discourse analysis. In *Discourse and dialogue*.

Van Wyk, E. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. Lexikos, 5, pp. 82-86.

Varantola, K. 2002. Use and Usability of Dictionaries: Common Sense and Context Sensibility? In M. Corréad (ed): Lexicography and Natural Language Processing: A Festshrift in Honour of B.T.S Atkins, pp. 30-44. Stuttgart: Eurolex.

Warburton, K., 2015. Managing terminology in commercial environments. Handbook of terminology, 1, pp.359-391.

Wardhaugh, R., 1972. Introduction to Linguistics.

Wiegand, H. 1983. On The Structure and Contents of a General Theory of Lexicography. Available at: www.eurolex.org/elx_proceedings/Eurolex1983/007_Herbert. [Accessed: 13 March 2014].

Wierzbicka, A., 1985. *Lexicography and conceptual analysis*. Ann Arbor [Mich.]: Karoma.

Xiao, R.Z., 2007. What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian JELT: Indonesian Journal of English Language Teaching*, *3*(1), pp.1-19.

Zgusta, L., 1971. Manual of Lexicography (Janua Linguarum, Series Maior 39) The Hague: Mouton.

Annexure 1: The Prototype Dictionary – Sehlalosi sa Sesotho sa Baithuti

a */a/ (sehlonngwanthao)* **1** sehlongwanthao sa mabitso a sehlopha sa pele: *matjhaba a utlwe sesotho melodi* **2** *(lehokedi)* lentswe le bontshang kamano dipakeng tsa sehlooho le leetsi polelong: *e se e le moshanyana a tsamayang ntle le ho itshwarella* **3** *(lehokedi la leemediqho)* lentswe le bontshang kamano pakeng tsa sehlooho le leemediqho polelong: *bao matla a puso a dutseng ho bona* **4** *(lehokedi la lerui)* lentswe le bontshang kamano pakeng tsa lebitso le lerui : *le ba tshireletso ka mafapha ohle a bona* **5** *(lehokedi la lesupi)* lentswe le hokelang lebitso ho se etswang: *mphemphe e a lapisa, bashweshwe* **6** *(lehokedi la leakaretsi)* lentswe le bontshang kamono pakeng tsa sehlooho le leakaretsi: *matswalla a lesotho* **7** *(lehokedi la leamanyi)* lentswe le hokelwang pele ho leamanyi: **8** *(lehokedi la lebadi)* lentswe le hokellwang pele ho palo: *wekshopo ena ya matsatsi a mararo e reretswe ho neha bankakarolo dintlhatheo.*

aba /a-ba/ *(leetsi)* **1** tsela ya ho fana ka dintho: *re tla aba mokudi meriana le ha a se e tswile sepetlele.*

abela */a-be-la/* **(=aba)** *(leetsi)* **1** ho fana ka karolo ya dintho tse itseng: *mmoloki o tla re abela tsona kajeno, a re thabiseka tshepo le ka tumelo.*

abelwa */a-be-lwa/* **(=aba)** *(leetsi)* **1** ho fuwa karolo ya dintho tse itseng: *kgau e tla abelwa baithuti ba dimasetase le bongaka selemo le selemo.*

abile */a-bi-le/* **(=aba)** *(leetsi)* tlhalosi ya hore ho itseng ho ile ha fumahantshwa nako e fetileng: *ehlile re di abile letsholong le fetileng.*

abuti */a-bu-ti/ (lebitso)* **1** ngwana wa moshanyana a moholo ho babang tlhahlamanong: *na banneng ba leloko la heno hona le ba bolotseng sepetlele (ntate, abuti, ba bang)?*

adima */a-di-ma/ (leetsi)* ho tshwarisa motho e mong phahlo ya hao : *o ne a atisa ho mo adima mohope ha ba ilo kga metsi.*

ahela */a-he-la/ (leetsi)* **1** ho etsetsa motho kapa ntho bodulo ka majwe kapa ditene: *mmuso o tla ntlafatsa bosebeletsi ba sepolesa sa Lesotho ka ho se ahela matlo a bodulo, diofisi, ho ba rupela le ho ba fumantsha thepa ya tshebetso.*

ahileng */a-hi-le-ng/ (leetsi)* **SHEBA ahile**

ahlame */a-hla-me/ (leetsi)* **1** ho bula molomo: *e ne e re ha a kentse e ngwe ya mese ya hae o mona o beyang dirotswana kantle bahlankana ba sale melomo e ahlame jwale ka ya marakana a bonang mmutla.*

ahloletswe */a-hlo-le-tswe/* **(=kahlolo)** *(leetsi)* **1** nehelano ya kahlolo bakeng sa tlolo ya moloa: *moqusuwa wa boraro moeketsi o se a ahloletswe bophelo bohle tjhankaneng kamora hore a hlalose a le molato diqosong tsohle tse neng di le kgahlano le yena.*

ahlolwa */a-hlo-lwa/* **(=kahlolo)** *(leetsi)* **1** ho fuwa kotlo ka lebaka la diphoso tseo motho a di entseng: *dimanusekripete di ahlolwa ntle le ho tsejwa ha mongodi.*

ahwa */a-hwa/* **(=aha)** *(leetsi)* **1** ketso ya ho bopa moaho **2** *(leetsi)* ho phahamiswa ka ho fuwa dikeletso kapa malebela: *metse e ahwa ke setjhaba.*

AIDS *(lebitso)* **1** kgotsufatso ya kwatsi ya bosolla tlhapi: *tlalehong ho itswe re hlalose hore re entseng ho fokotsa sekgahla sa HIV le AIDS le ho thusa dikgutsana*. [English]

ajwa */a-jwa/ (leetsi)* **SHEBA aba**: *e tla ajwa hape ka mora dilemo tse ding tse pedi.*

akanya */a-ka-nya/ (leetsi)* ho ba le mohopolo ka ho itseng: *ke akanya hore ho ka ba molemo ha re ka ra tsohella teng ka matjeke.*

akarelleditsweng /a-ka-re-lle-di-tswe-ng/*(lekgethi)* sephetho se fumanweng ha ho etswa qeto, se kenyelleditseng tsohle: *pele o ka nka karolo, ho batleha hore o utlwisise lebaka leo re etsang sena le dintho tse akarelleditsweng.*

akarelletsa /a-ka-re-lle-tsa/ *(lekgethi)* **1** ho sheba dintlha ka tsela e le nngwe: *boitsebiso ba hao bo leng ditokomaneng, bo jwalo ka dilemo tsa hao, histori ya hao ya tsa maphelo ho akarelletsa le meriana e meng yeo o ka e nwang.*

akaretsa */a-ka-re-tsa/ (lekgethi)* **1** ho kenyeletsa dintlha ka tlasa ntho e le nngwe: *o se ke wa sebedisa dithethefatsi tse ding kapa meriana ho akaretsa le meriana ya dingaka tsa sesotho* **2** *(lekgethi)* ho aparela ka bophara le ho nama: *thuso ya tjhelete e tla fuwa baahi bohle ho akaretsa le melata.*

akaretsang /a-ka-re-tsa-ng/ *(lekgethi)* **SHEBA akaretsa**: *tshebediso ya lereo le akaretsang moelelo wa diphetolelo le botoloki*

akere */a-ke-re/ (lekopanyi)* lentswe le sebediswang ho nnetefatsa le ho toboketsa taba: *hao, akere re sa bua.*

akga */a-kga/ (leetsi)* ho lahla matsoho ka tsela ya ngala: *a akga matsoho a tsamaya ha a bona hore ha ba mo nkelle hloohong.*

akgang */a-kga-ng/ (leetsi)* **SHEBA akga**: *le se ke la akgang matsoho ha le kgenne hoba le pele di na le baji.*

akgela */a-kge-la/ (leetsi)* ho lahlela ntho hodimo kapa ho e kenya ka hare ha ho hong: *ba kgothaletswa ho akgela ditshisinyo tsa bona ka kakaretso ka Senyesemane le Sepenishe.*

akgetse */a-kge-tse/ (leetsi)* **SHEBA akga**: *dihloowana ba di akgetse morao.*

akile */a-ki-le/* **(=aka)** *(leetsi)* ho tshwara motho ka tsela ya ho mo aka: *o ne a se a akile mmae o bona hore o tla mo hlolohela.*

akofang */a-ko-fa-ng/* **(=akofa)** *(leetsi)* taelo ya hore dintho di etswe kapele: *batjha ye, akofang!*

ala */a-la/ (leetsi)* ho beha dintho ka mokgwa o pharalletseng: *ba fihlile hang ha a sa tswa ala meseme.*

alafa */a-la-fa/ (leetsi)* ho nka karolo ya booki ho thusa motho ya kulang: *inshorense ya bathusi e tla lefella meriana ya ho alafa ho kula ha hao kapa dintho tse bakilweng tshebetso ena.*

alafeha */a-la-fe-ha/* **(=alafa)** *(lehlalosi)* ha ho na le kgonahalo ya ho fola: *seo se buoa hore HIV ha e sa alafeha ka meriana e meng ya hiv, seo se bolelahore meriana e ka se sebetse ka tsela yeo ho nahanoeng ka yona.*

alame */a-la-me/ (leetsi)* ha kgoho di dula hodima mahe ho a futhumatsa hore a tle a qhotse ditsuonyana: *enore ka thapama a di kokose a di lebise phaposing ya mae hore di alame.*

aletare */a-le-ta-re/ (lebitso)* sebaka ka hare ho ntlo ya thapelo moo ho rapelwang teng kapa ho dulang baholo ba mokgatlo wa bodumedi: *ke thabile ka aha aletare – ya bethele*. [English]

alola */a-lo-la/ (leetsi)* ho hleka diphate kapa sebaka seo ho neng ho robetswe ho sona: *a tsoha a alola mme a*

itlhophisetsa leeto.

alosa */a-lo-sa/ (leetsi)* ho disa diphoofolo: *o hopole hore bana ba ka kgale ba alosa dintho tsena mme ka lebaka leo tsa ka di tlameha ho ba ngata ho feta tsa Botle.*

ama /a-ma/ *(leetsi)* ho thetsa: *phethoho ena e ama dimasepala tseno tse tharo.*

amana */a-ma-na/ (leetsi)* **1** ho thetsana: *o ne a iphethela feela yaba ba amana.* **2** *(lekgethi)* ho ba le dikamano tse itseng: *ha ho na etswa ditekolo bakeng sa maemo a itseng ntle le ha eba maemo ao a amana le patlisiso.*

ambulense */a-mbu-len-se/ (lebitso)* koloi e tsamaisang bakudi ba hlolehang haholo: *bitsang ambulense bo! O idibane.* [English]

amehile /*a-me-hi-le/ (lekgethi)* ho tsietswa ke ditlamorao tsa ho itseng: *tabeng ena, lesotho le lona le amehile haholo.*

amen! */a-men /* ha ho be jwalo! ke lentswe le kwalang thapelo ya bakereste.: *phutheho e ile ya re **amen** ha moruti a qeta ho rapela.* [English]

Amerika */a-me-ri-ka/ (lebitso)* kontinente; e arotswe ka leboya le borwa: *boemo ba loti ho dolara ya Amerika le bona bo thatafaditse ditaba hoba theko ya thepa kantle ho naha, re nkella le tsona dijothollo, e nyolohile haholo.*

amoha */a-mo-ha/ (leetsi)* ho nka ntho hampe mothong e mong: *dipina tsa mokola, matshwele di a amoha menahano.*

amohela /*a-mo-he-la/ (leetsi)* **1** ho kgola moputso: *le yena o ne a amohela tjhelete ya boqeku*: **2** ho dumella baeti sebakeng: *o tla ntlafatsa dirapa ka ho aha matlo a ho amohela baeti le ditsi tsa boithuto.*

amohelang */a-mo-he-la-ng/ (leetsi)* **SHEBA amohela**

ananela */a-na-ne-la/ (leetsi)* ho bontsha kutlwisiso ya seo ho buuwang ka sona: *ditaba tsena Botle le yena a di ananela.*

anathema */a-na-the-ma/ (lebitso)* thohako e lebiswa ho motho ha a entse ketso e sa tshwareleheng **:** *ha se jwale ke mo kgalemela, jwale ho lekane, ke mo fa anathema!* [Sekgeriki]

anela */a-ne-la/ (leetsi)* ha ntho e phatlalletse: *dingaka hoo di ka ba anela kaofela.*

anetse /a-ne-tse/ *(leetsi)* ha dintho di lekanetse ka tsela e hlokehang: *mme re tla ithuta ka nako ya selemo hore na di anetse tshebetso jwang naha ho pota.*

ao! */a-o/ (lelahlelwa)* **1** lentswe le bontshang makaalo moqoqong: *ao! le wena o ne o ile?* **2** *(lehokedi)* lentswe le hokelang sehlooho le ketso polelong: *maqheka ao o a entseng a tla o tswela kgomo.*

apara /*a-pa-ra/ (leetsi)* ho beha diphahlo hodima mmele: *le hoja he a apara marantha, ke motho a itlhomphang.*

aparetse */a-pa-re-tse/ (leetsi)* ho apara diphahlo ka sepheo sa ho supa hohong kapa ka lebaka la ketsahalo e itseng: *re aparetse letjhato.*

api /a-pi/ *(lebitso)* mofuta wa tshwene e monyane e masene haholo: *ka ntlong ya hae ho thwe o ne a ruile api.*

apoha */a-po-ha/ (lehlalosi)* ho hlaka ka maemo a lehodimo: *pula e emisitse ebile maru a apoha*.

araba */a-ra-ba/ (leetsi)* ho fa motho karabo ya dipotso tsa hae: *o fana ka lesedi le lekaneng ho araba dipotso tsa ngwana ka botlalo.*

arabiswa /*a-ra-bi-swa/ (leetsi)* ha motho e mong a bua kgahlano le se buuweng pele leha ho ne ho sa lebellwa hore a arabe: *o ne a sa dumela hore o arabiswa ke ngwana eo a mo tswetseng.*

arohana */a-ro-ha-na/ (leetsi)* ho hlalana kapa ho siyana ha batho kapa dintho: *Mokopu le Baratang ba tsamaile mmoho yare ha ba fihla nokeng ya mmasemouse ba arohana.*

arola */a-ro-la/ (leetsi)* ho beha ntho e neng e kopane le e nngwe ka thoko ho ya bo yona: *hoba ke qete tsena kaofela ke tla theola dintho tsohle tse thabeng ke tlo di arola.*

arubeditse /*a-ru-be-di-tse/ (leetsi)* ho kwahela ka morero wa ho kwalla motjheso ka hare bakeng sa ho futha: *o sa arubeditse hobane o thibane dinko ke sefuba.*

ata */a-ta/ (leetsi)* ha ntho e phatlalletse, e tsebahala ka ho phara: *mapolasing botlokotsebe ha bo a ata jwaloka metsesetoropong.*

atamela */a-ta-me-la/ (leetsi)* ho tla haufi le ntho kapa motho: *a botsa ka kgalefo e kgolo, e le ha a ntse a atamela tjhebeleng moo lelapa le neng le dutse.*

aterese */a-te-re-se/ (lebitso)* sebaka seo motho kapa mokgatlo o hlomilweng ho sona; ha ngata sebaka sena se na le nomoro e leng e tsamayang le le lebitso la seterata le motse setoropo: *haeba o rata ho kenngwa lenaneng la ho romellwa ditsebiso, romela lebitso la hao ka imeili, mokgatlo oo o leng setho sa ona (haeba ho lokela) le aterese e ka sebediswang ho o romelang mangolo.* [English]

athe /*a-the/ (lekopanyi)* **1** lentswe le sebediswang ha motho a hloka tlhakisetso: *athe hase kgomo ke tladi!* **2** *(lelahlelwa)* lentswe le bontshang ho makala: *athe batho ha re tsebe hore nyewe e se e le teng ya majantja ba tseka hore moteaphala o ile a ja nako hobane pele ho phomolo o ile a ja metsotoso e mehlano.*

athebane /*a-the-ba-ne/ (lekopanyi)* lentswe le sebediswang ho atolosa tlhakisetso: *athebane re batla hore o itlhonepjwe.*

atikele */a-ti-ke-le/ (lebitso)* sephepheshana se ngotsweng mabapi le taba e itseng se fumanwang kahara bukana tsa phahlalatso tse kang dikoranta.: *ha o batla ho nehelana ka atikele mabapi le kgatiso tse tlang, re tsebise.* [English]

atjhe! /a-tjhe/ *(lelahlelwa)* lentswe le sebediswang ha motho a tenehile kapa a kgathetse matla a bile a inehela : *atjhe! ke buile ho lekane.*

atleha */a-tle-ha/ (leetsi)* ho fihlella ditabatabelo tsa hao: *setjhaba sa Basotho se ka be se sa atleha ha e ne e se ka tshebedisanommoho e matla, ya boitelo hape e mafolofolo e entsweng ke komiti ya naha, ya tjhebelopele.*

atlehile */a-tle-hi-le/* **(=atleha)** *(leetsi)* ha ho fihlelletswe sephetho se kgotsofatsang: *baratang a phakisa a elellwa hore maraba a hae a atlehile.*

atisa */a-ti-sa/ (leetsi)* **SHEBA ata** ho eketsa se teng hore se be sengata: *a tshela metsi ho atisa moritelo wa jwala boo a neng a bo phehetse banna bana ba hosele.*

ba */ba/ (lehokedi)* lentswe polelong le amanyang sehlooho sa sehlopha sa mabitso sa bobedi le leetsi, lehlalosi kapa lekgethi: *ke batho ba bangata ba ka mo emang nokeng ha ho se ho ka senyeha.*

baahi */ba-a-hi/* **(=aha)** *(lebitso)* batho ba phelang tulong e itseng:*mme ho fihlela ha jwale baahi ba re ha ho motho ya seng a ithaopile ho nka bokhanselara ba setereke seo.*

baahlodi */ba-a-hlo-di/ (lebitso)* batho ba etsang diqeto: *dinehelano di kwallwa ke moifo wa baahlodi ba tlhodisano.*

baba /*ba-ba/ (lehlalosi)* tatso ya dijo e tjhesang leleme: *tjhilisi e a baba*.

bababi */ba-ba-bi/ (lebitso)* batho ba opelwang mmeleng: *se tshwereng bababi bana se palletse methokgo.*

babadisisi /*ba-ba-di-si-si/ (lebitso)* batho ba hlwayang diphoso ditokomaneng: *maemo a hoba setho a butswe bakeng sa bafetoledi, bahlalosi, balekodi, babadisisi.*

baballa /*ba-ba-lla/ (leetsi)* ho beha tlasa tlhokomelo: *ho baballa tikoloho.*

babatledi */*ba-ba-tle-di/*(lebitso)* batho ba thusang ho fumana ntho kapa motho: *mokgatlo ona ha se wa babatledi ba mesebetsi.*

babatli */ba-ba-tli/ (leetsi)* batho ba batlang ntho kapa motho: *bafetoledi ha ba e patale ha ba thola mesebetsi le babatli ba ditshebeletso ha ba sebedise moralo wa rona wa tlhaisoleseding ho patala.*

babedi /*ba-be-di/ (lekgethollilebadi)* palo e latelang bonngwe: *kahlolo kgahlano le baqusuwa ba babedi ba setseng nyeweng ya polao e etsahetseng.*

babelaelwa */ba-be-la-e-lwa/ (lebitso)* batho bao ho nahanwang hore ba tlotse molao o itseng: *o re babelaelwa esale ba le diphuputso bakeng sa dikgwedi tse hlano.*

babohi */ba-bo-hi/ (lebitso)* batho ba shebelletseng ketsahalo: *ho tea mohlala, babohi ba matjhaba ha ba lemoha hore o ne a etsa botoloki bo sa nepahalang hoba ba sa tsebe puo ya mona.*

babolai */ba-bo-la-i/ (lebitso)* batho ba nkang maphelo kapa ba a fedisang: *empa tsa lefu la hae tsona wa di tseba, o tseba le babolai ba hae.*

babolawa /*ba-bo-la-wa/ (lebitso)* batho bao maphelo a bona a fedisitsweng: *ho fumanwe babolawa ba babedi feela.*

babosisi */ba-bo-si-si/ (lebitso)* batho ba tsamaisang molao: *ba tswalletswe boreneng ke ka hoo e leng babosisi.*

babotsi */ba-bo-tsi/ (lebitso)* batho ba botsang dipotso dipuisanong: *o tla lebellwa ho fa babotso lesedi leo ba le batlang.*

babuelli */ba-bu-e-lli/ (lebitso)* batho ba emelang batho dinyeweng tsa bona: *dipatlisiso tsena di ne di tswa bathong ba neng ba hloka diphetolelo ka mabaka a botho, le mekgatlong ya thuto, babuelli le dikgwebo tse fapafapaneng.*

babusi /*ba-bu-si/ (lebitso)* batho ba etelletseng pele, bao matla a puso a dutseng ho bona: *babusi ba kopane mmoho.*

badimo /*ba-di-mo*/ **(=modimo)** *(lebitso)* batho ba hlokahetseng bao ho dumelwang hore ba sireletsa beng ka bona ba setseng lefatsheng: *badimo ba lepoqo bane ba sa tsebe ditaba tsa kereke.*

badireng /*ba-di-re-ng*/ *(lebitso)* batho ba seng lehlakoreng la hao, ba o lakaletsang bobe: *motse o tjhesitswe ke badireng.*

baditjhaba /*ba-di-tjha-ba*/ **(=setjhaba)** *(lebitso)* batho ba matjhaba a kantle ho setjhaba sa heno: *etjhong, he, Basotho, amohelang baditjhaba.*

baetsana /*ba-e-tsa-na*/ *(lebitso)* metswalle le ditho tsa leloko tse tshehetsang monyaduwa ka letsatsi la lenyalo la moetlo wa Sekreste: *o kopile nna le kgaistedi ya hae hore re be baetsana mohla letjhato la hae.*

baferekanyi */ba-fe-re-ka-nyi/* **(=ferekanya)** *(lebitso)* batho ba senyang kapa ba qabanyang; ba bakang pherekano: *baferekanyi ba ntshetse marao, bomopheme ba ntlola bosiu.*

bahahlaudi */ba-ha-hla-u-di/ (lebitso)* batho ba tsamayang lefatshe ho ithuta kapa ho ithabisang: *moo teng batlang ho rekisa jwalo ka sebokeng se enang le bokgoni ba ho nonya bahahlaudi.*

bahale */ba-ha-le/ (lebitso)* batho ba bileng le seabo tokolohong ya setjhaba, ba lwaneng dintwa molemong wa bohle: *bahale ba ba kgabola morokotlo.*

bahetene */ba-he-te-ne/ (lebitso)* tumelong ya Sekreste, ke batho ba sa pheleng ho latelang molao wa Sekreste, ba etsang ketso disele tse thunthetsang seo tumelo ena e se kgothalletsang: *ke bahetene ba senyang thero ya modimo.* [English]

bahiri */ba-hi-ri/ (lebitso)* batho ba fanang ka mesebetsi: *tsa kamora tlohello ya mosebetsi tse tla lefuwang ke bahiri/borramesebetsi hore e be tlwaelo*.

bahlabani */ba-hla-ba-ni/* **(=hlaba)** batho ba leng ntweng: *tladi ya mofuta ya hobela hara bahlabani - ke kgomo, selo, namane ya sekgutlo.*

bahlahisi */ba-hla-hi-si/ (lebitso)* batho ba thomileng ntho e itseng: *feme ya maqwetha a bahlahisi ba ditaba, mokgatlo.*

bahlahlobi */ba-hla-hlo-bi/ (lebitso)* batho ba nnetefatsang ho nepahala kapa ho fosahala ha ditaba: *bahlahlobi ba dingolwa ba sebetsang moralong wa tsa puo.*

bahlahlobuwa */ba-hla-hlo-bu-wa/ (lebitso)* batho ba ngolang tlhalobo: *ka tsela e tshwanang, teko ya boikwetliso e raletswe ho sesefa boitokisetso ba bahlahlobuwa bakeng sa nnetefatso boiphihlelong, kapa boitokisetsong tlhahlobo.*

bahlalefi */ba-hla-le-fi/ (lebitso)* ditsebi tsa ditaba tse itseng: *ho tla botswa bahlalefi ho nnetefatsa hore dintlha di nepahetse.*

bahlalosi */ba-hla-lo-si/ (lebitso)* batho ba manollang ditaba: *ha ho phapang pakeng tsa ditoloko le bahlalosi.*

bahlanka */ba-hla-nka/ (lebitso)* batho ba sebeletsang setheo se itseng: *bahlomphehi ditho tse kgabane tsa matlo a mabedi a palamente bahlanka bohle ba setjhaba bomme le bontate.*

bahlankana */ba-hla-nka-na/ (lebitso)* bashanyana ba dilemong tse hodimo: *ke ditaba dife tse o di utlwileng, tse ntle kapa tse mpe, ka bahlankana ba dilemong tsa hao ba bolotsweng ke mofani wa tlhokomelo ya boitekanelo?*

bahlodi */ba-hlo-di/ (lebitso)* batho ba hapileng kapa ba hlotseng hohong: *bahlodi ba fetileng ba moputso ona ba kenyeletsa baithuti ba bararo.*

bahloki /ba-hlo-ki/ **=hloka** *(lebitso)* batho ba futsanehileng: *haesale e le tshebetsong bakeng sa dilemo tse fetang tse 50 mme ke lebala la pele leo bahloki ba ditshebeletso ba tlang ho lona ho fumana ditshebeletso.*

bahlokomedi */ba-hlo-ko-me-di/ (lebitso)* batho ba hlokomelang dintho: *haeba o kula ka mora patlisiso ena, bahlokomedi ba hao ba tsa maphelo, (dingaka, baoki, baeletsi) ba ka batla ho ikgokahanya le batho ba etsang patlisiso ena.*

bahlomphehi */ba-hlo-mphe-hi/* **(=tlhompho)** *(lebitso)* batho ba nang le maemo a holimo sechabeng mme ba hlompheha: *a sechaba mohlomphehi moahloli e moholo mohlomphehi motlatsi oa tona-kholo bahlomphehi matona a khabane bahlomphehi baahloli ba lekhotla le phahameng bahlomphehi baemeli ba linaha tse nang le setsoalle le lesotho, le baemeli ba mekhatlo ea lefatše.*

bahlophehi /*ba-hlo-phe-hi*/ *(lebitso)* batho ba hlokang ho iphedisa: *ke hlahile ke le mothusi wa bahlophehi, mosebetsi wa ba bolawang.*

bahlophisi */ba-hlo-phi-si/ (lebitso)* batho ba tshepetsweng ho hlophisa diketsehalo: *wekshopo ena ya semetletsa e tla thusa bahlophisi kaofela ba batlang ho ithuta ka kapa ka ho itjhorisa ditsebo tsa bona tsa puo.*

bahlwekisi /*ba-hlwe-ki-si*/ *(lebitso)* batho ba hlwekisang dibaka: *bahlwekisi ba sepetlele, le ba bang, ba entse hore o ikutlwe o phuthulohile?*

bahodisi /*ba-ho-di-si*/ *(lebitso)* batho ba elang hloko kgolo ya bana: *ke boetse ke etsa boipiletso ho batswadi le bahodisi ba barwetsana bao ba Basotho.*

baholo /*ba-ho-lo*/ *(lebitso)* batho ba okametseng ba bang: *baholo ba makala ke bona ba nang le matla a ho nka diqeto.*

baholwane /*ba-ho-lwa-ne*/ *(lebitso)* bo abuti ba baholo: *o na le baholwane ba bollotseng?*

bahwebi /*ba-hwe-bi*/ *(lebitso)* batho ba rekisang: *batho bana e ne e le bahwebi ba dithetefatsi.*

baikemedi /*ba-i-ke-me-di*/ *(lebitso)* batho ba ikemetseng ka hohlehohle: *ho tla ba le dithupelo tse tla ba molemo haholo ho baikemedi.*

baikgwantlelli /*ba-i-kgwa-ntle-lli*/ *(lebitso)* batho ba itshebeletsang: *ho tla ba le dikopano tse robong kaofela tse tla kgahla baikgwantlelli.*

baikopanyi /*ba-i-ko-pa-nyi*/ *(lebitso)* batho ba ikopantseng ho fihlela sepheo se le seng: *mmuso o tla kenya tshebetsong leano la ho thusa bahwebi ba banyenyane le baikopanyi ka ho ba rupella le ho ba thusa ho kadima ditjhelete ho ithehela mesebetsi.*

bainehedi */ba-i-ne-he-di/ (lebitso)* batho ba itlamang ho latela tsela e itseng: *o ka re bainehedi ba bakae ba amohetseng thomo ya bona?*

baiphihledi /*ba-i-phi-hle-di*/ *(lebitso)* batho ba nang le tsebo e batsi ka dithuto tse itseng tsa boemo bo hodimo: *boholo ke tsa baiphihledi ba yunivesithi, empa ho na le diwekeshopo tse ding.*

baithaopi /*ba-i-tha-o-pi*/ *(lebitso)* batho ba tswang sehlabelo ho thusa ka tshebeletso e itseng ntle le tefo: *phetolelo e etswa ke baithaopi ka nako tse fapaneng.*

baithuti /*ba-i-thu-ti*/ *(lebitso)* batho ba baholo ba kenang dikolong tse phahameng tse kang diyunivesithi kapa dikholeiji: *baithuti ha ba sa tla hlola ba kgona ho ngola teko ya boikwetliso.*

bajalefa /*ba-ja-le-fa*/ **(=ja + lefa)** *(lebitso, lebitsokgoboka)* batho ba sielwang maruo ke beng ka bona: *ke tokomane e bohlokwa hoba e tla thusa balaodi ba mokga ha ba neha bajalefa bao o ba kegthile ditjhelete.*

baka /*ba-ka*/ *(leetsi)* **1** ho susumetsa hore ho be le se etsahalang: *hona ho baka morwalo o sa hlokahaleng tsamaisong ya mokgwa wa tlhahlobo.* **2** *(lehlalosi)* kgutsofatso ya lentswe sebaka: *kenang bohle kenang bohle, baka se sa le teng. (lekopanyi)* kgutsofatso ya lentswe lebaka: *bana ha Mokopu ba dieha haholo ha tshwara ntata bona tseleng, ka baka la ho totoba ha mma bona.*

bakela /*ba-ke-la*/ *(leetsi)* ho tlisetsa motho sesosa sa ntho e itseng: *a bona hore melato ya hae e kentse mmae mathatheng le ho mo bakela tsietsi e kgolo.*

bakenedi /*ba-ke-ne-di*/ *(lebitso)* batho ba bang le seabo diketsahalong: *bakenedi ba tsheletseng ba kenyeleditswe mokgahlelong ka mong.*

baketse /*ba-ke-tse*/ *(leetsi)* **BAPISA bakela**

bakganni /*ba-kga-nni*/ **(=kganna)** *(lebitso)* batho ba laolang kapa ba tsamaisang dipalangwang: *mohwanto wa bakganni ba dibese o siya bapalami ba eme le monnamateneng dibeke tse tharo kaofela.*

bakgethi /*ba-kge-thi*/ *(lebitso)* batho ba hlophang maetedipele ba dipolotiki: *puso ya setjhaba ka setjhaba e matlafatswa ke ha bakgethi ba etswa ka bongata ho kgetha baemedi ba bona.*

bakgotsi /*ba-kgo-tsi*/ *(lebitso)* batho ba nang le dikamano tsa setswalle: *bakgotsi ba hao ba tla nahana eng ha o ka bolotswa?*

baki /ba-ki/ *(lebitso)* **1** mofuta wa koloi e nyane eo hangata e sebedisetswang ho laetjha thepa. **2** *(lebitso)* seaparo sa ka hodimo se kang jase empa sele bobebe ho feta jase: *ntate o apara baki ha a ya kerekeng.*

bakone */ba-ko-ne/ (lebitso)* batho ahileng karolong e bokone ya sebaka: *lakabe la tlola thoteng ya bale, dibataolong, athe le qhaqholla diteng tsa bakone.*

bakopani */ba-ko-pa-ni/ (lebitso)* batho ba ikopantseng ho fihlela sepheo sele seng: *bakopani thite ya kgathwa ke lesotho thoteng botswana le swatsing tsa e kgobola manyolo le metsi ho etsa katleho.*

bakreste */ba-kre-ste/ (lebitso)* batho ba latelang tumelo ya sekreste: *bakreste ka mafapha ohle le ditumelo tsa bona ba keteka tswalo ya lesea Jesu kreste le thakangwaha ho yena.*

bakudi */ba-ku-di/ (lebitso)* batho ba opelwang mmeleng; ba sa ikutlweng hantle kelellong kapa moyeng: *dithuto tsa patlisiso di etswa ho fumana ditsela tse nepahetseng tsa ho alafa bakudi.*

bakwa */ba-kwa/ (leetsi)* **BAPISA baka**

bakwena */ba-kwe-na/* **(=kwena)** *(lebitso)* batho ba anang kwena: *ke kgale le kgotsa ho lapa, bakwena; ho mengwaha le supaka ditlhobolo.*

bakwenehi */ba-kwe-ni-hi/ (lebitso)* batho ba fapohileng mererong eo ba ileng ba itella ona: *bongata bo hlome ke la ho qetela, le bakwenehi ba hopole thapelo!*

bala */ba-la/ (leetsi)* ho sheba se itseng ho tswa dingolweng: *ha taba di le jwalo mantsiboya a nna a ba kenya sekolo a ba ruta ho bala le ho ngola.*

balaudi */ba-la-u-di/ (lebitso)* batho ba tshepetsweng ho laola tshebetso: *balaudi ba tsa bohahlaudi Free State, ba ikakgele ka setotswana tsamaisong ya baeti.*

balebelli */ba-le-be-lli/ (lebitso)* **1** batho ba shebelletseng ketsahalo: *ke lakatsa ho leboha balebelli ba dikgetho.* **2** *(lebitso)* batho ba disang sebaka: *re dumela hore ka thuso ya balabelli, re ka kgolwa hore mebaraka ya rona e sireletsehile.*

balefapha */ba-le-fa-pha/ (lebitso)* batho ba sebetsang lefapheng le itseng: *re lakatsa ho ithuta ka karolo ya balefapha la tsa bophelo bo botle ho fa batjha ditshebeletso tsena.*

baleha /*ba-le-ha*/ *(leetsi)* ho raoha ka lebelo: *etswe motho e mobe o a baleha, leha a sa phallelwe.*

balekani */ba-le-ka-ni/ (lebitso)* batho ba nang le tumellano ya kamono e itseng *tlhokeho ya ho fokotsa palo ya balekani ba thobalano.*

balekodi */ba-le-ko-di/ (lebitso)* batho ba hlolang maemo: *balekodi ba tla sekaseka kopo ya hao o bona hore na dintlha tsohle di nepahetse.*

balellwa */ba-le-llwa/ (leetsi)* ho akaretswa palong e itseng: *e balellwa hara makgabane a hlokehang bakeng sa tshebetsontle.*

baleloko */ba-le-lo-ko/ (lebitso)* ditho tsa lelapa le le leng: *baleloko la hao le setshaba ba ikutlwa jwang ka qeto ya hao ya ho batla hore mora wa hao a bolotswe sepetlele?*

balemi /ba-le-mi/ *(lebitso)* batho ba iphedisang ka temo: *ka letsatsi le leng Botle a roma ngwana wa Baratang e monyenyane masimong hore a isetse balemi dijo.*

balemohi */ba-le-mo-hi/ (lebitso)* batho ba nang le tjhadimeho ka ditaba: *e bile matsatsi a mararo a mosebetsi o mongata, mme ditho tse nne tsa sati di ile tsa kenela dikarolo tsa kopano e le balemohi.*

balepera */ba-le-pe-ra/ (lebitso)* ka Sekreste, ke batho ba bao dibe tsa bona di ba bakelang mahloko a letlalo: *eitse ha a kena motsaneng o mong, a kgahlana le banna ba leshome, e leng balepera.* [English]

balepi */ba-le-pi/* **(=lepa)** *(lebitso)* batho ba akanyang diketsahalo tsa bokamoso: *re sentse ka ho latela keletso ya balepi ba maemo a lehodimo.*

balwetsi */ba-lwe-tsi/ (lebitso)* **LETSHWANI bakudi**

baloi */ba-lo-i/ (lebitso)* batho ba loyang: *ke bao he baloi ba re kgathatsang, re lokisetse ditaba ka potlako.*

balopollwa */ba-lo-po-llwa/ (lebitso)* batho bao ho thweng ba tswetswe kgetlo la bobedi mme ba na le tumelo e tiileng ho modimo: *naha e teng naha e teng e ratehahang e kenwang ka balopollwa*.

balotsana */ba-lo-tsa-na/ (lebitso)* batho ba sa tshepahaleng: *mahlo a le bohale, a le nchocho, bile a hlaba, ho bonahala hantle hore ke ea tloaetseng ho sheba lichaba-chaba tsa batho le ho khetha balotsana ho ba lokileng.*

bamamedi */ba-ma-me-di/ (lebitso)* batho ba mamelang sealemoya, mmino kapa dipuisano ka marangrang a

fapaneng: *bina ka nko kgeleke tsa kwano di bina ka kgitla ho se bamamedi, di bina sekene.*

bamemuwa */ba-me-mu-wa/ (lebitso)* batho ba bileditsweng sebakeng: *bamemuwa ba subuhlellana moketeng wa habo.*

banana */ba-na-na/ (lebitso)* bana ba barwetsana: *o nahana hore banana ba dilemo tsa hao ba nahana eng ka bashanyana ba dilemong tsa hao ba bolotsang?*

bane */ba-ne/ (lekgotholli)* ha palo ya batho e le nne: *ba bane bona ba ne basa tswa dikotsi.*

banehedi /*ba-ne-he-di/ (lebitso)* batho ba fanang ka dintho setjhabeng: *banehedi ba mmisa ba tla etela malapeng.*

banehelwaditshebeletso /*ba-ne-he-lwa-di-tshe-be-le-tso/ (lebitso)* batho ba fuwang ditshebeletso: *ke banehelwaditshebeletso ba ba kae bao o ba nehang tshehetso letsatsing le tlwaelehileng?*

banewa *(lebitso)* **SINONIME banehelwa**

bangata */ba-nga-ta/ (lehlalosi)* palo e hodimo ya batho: *ho ne ho sena batho ba bangata mmileng.*

bangodi /*ba-ngo-di/ (lebitso)* batho ba hatisang ditaba ka ho di ngola fatshe: *dintlha ka botlalo di tla tsejwa ke bangodi.*

baninyana */ba-ni-nya-na/* **LETSWHANI banana**

banka */ba-nka/ (lebitso)* setheo sa poloko ya tjhelete: *mmuso o tla tswela pele ho sebedisana le banka e kgolo ya lefatshe.*

bankakarolo */ba-nka-ka-ro-lo/ (lebitso)* batho ba nang le seabo ketsahalong: *wekshopo ena ya matsatsi a mararo e e reretswe ho neha bankakarolo dintlhatheo tsa mosebetsi wa mofuta ona.*

baphatlalatsi */ba-pha-tla-la-tsi/ (lebitso)* mekgatlo e hlahisang le ho pepesa ditaba ekaba dingolwa, mmino kapa mananeo a marangrang: *baphatlalatsi ba ikemetseng ba ananelwa ka mosebetsi o bohlokwa oo ba o etsang setjhabeng.*

bapisana */ba-pi-sa-na/ (leetsi)* ho ba malebana kapa ka thoko ho se seng: *iketsetse dintho tse kgannehang tse bobebe kapa tse kgonang ho bapisana.*

bapiswa */ba-pi-swa/ (leetsi)* tekolo ya ntho kgahlano le e nngwe: *pilisi eo e bapiswa le pilisi e nngwe eo le yona e kopantsweng ya HIV.*

baputswa */ba-pu-tswa/ (lehlalosi)* ho hlalosa batho ba ditshila: *kgomo tsa batho ba baputswa ho ata di ata jwalo ka moriri wa hlooho motho wa marumo ha epelwe hae, motho wa marumo o epelwa naheng.*

baqapi */ba-qa-pi/ (lebitso)* batho ba qapang ditaba: *hona ho bolela hore ha nako e ntse e eya baqapi ba mareo ba tla lokela ho ngodiswa le ho ba le mangolo tlhomamiso ele hore ba sebetse.*

baqhekanyetsi */ba-qhe-ka-nye-tsi/ (lebitso)* batho ba unang melemo ka diketso tsa leshano: *tokomane eo ya ho eta e ne e sebeliswa hampe ke baqhekanyetsi dibankeng le dikgwebong ka hara naha.*

baqolotsi /*ba-qo-lo-tsi/ (lebitso)* batho ba utullang ditaba ka morero wa ho di tlaleha: *tebello ke hore letona la sepolesa, mebila le dipalangwang, a tshetlehele baqolotsi ba ditaba puo tebang le marangrang a mehato e tlang*

ho kenngwa tshebetsong ya tshireletso mebileng.

baqusuwa */ba-qu-su-wa/ (lebitso)* batho ba fumanweng molato: *kahlolo kgahlano le baqusuwa ba babedi ba setseng nyeweng ya polao e etsahetseng motseng wa Qwaqwa.*

bara */ba-ra/ (lebitso)* bana ba motho ba batona: *bara le bona ba amohela pina ena ea mma bona.*

baradi */ba-ra-di/ (lebitso)* bana ba batshehadi: *empa o ile a bona ona e le monyetla wa ho ruta baradi ba hae ba tle ba ithlokomele.*

bararo */ba-ra-ro/ (lekgetholli)* ha palo ya batho e fihla ho tharo: *ka hare ho yona e ne e le banana ba bararo ba bonahala eka ha ba thaba.*

baratuwa /*ba-ra-tu-wa/ (lebitso)* batho ba ratwang haholo: *wena le baratuwa ba hao le keteke ha monate matsatsing ana a phomolo.*

bareki */ba-re-ki/ (lebitso)* batho ba patalang tjhelete ho fumane thepa kapa ditshebeletso: *phetolelo ya setekgeniki le mareo le taolo ya ditebello tsa bareki ba tshebeletso ena.*

barekibafepedi */ba-re-ki-ba-fe-pe-di/ (lebitso)* borakgwebo ba rekang thepa ka morero wa ho e rekisetsa ba batho bang: *mekgwa e metjha eo barekibafepedi ba e sebedisang ho phalella bareki ba bona e sebedisa marangrang.*

baromuwa */ba-ro-mu-wa/ (lebitso)* batho ba tshepelwang ho fetisa melaetsa: *ka kgwedi ya Phato ke baromuwa baleshome ba tswileng letsholo ho phallela setjhaba.*

barui */ba-ru-i/ (lebitso)* batho ba nang le matlotlo: *boitsebiso bo fumanweng bo ka romelwa ho ditho tse ding tsa dikhampani tsa barui.*

barupelli */ba-ru-pe-lli/ (lebitso)* batho ba nang le tsebo ya ho ruta ba bang: *ke motlotlo ho ba e mong wa barupelli lenaneong lena la ho ntshetsa batjha pele.*

barutehi */ba-ru-te-hi/ (lebitso)* batho ba nang le tsebo e hodimo kapa e ngata: *athe le hlalefetseng, barutehi?*

baruti */ba-ru-ti/ (lebitso)* batho fetisetsang tsebo kapa lesedi ho ba bang: *baruti ba Moshoeshoe ba tshepisitse ho hasa efakgedi setjhabeng sa Basotho ho se fahlolla.*

barutise */ba-ru-ti-se/ (lebitso)* **SINONIME barupelli**

Barwa */ba-rwa/ (lebitso)* batho ba mohlobo wa Makhoi bao ho neng ho tsejwa ba phela naheng, ba phelang ka ho tsoma le ka meroho ya naheng: *mekgwa ya hae ekare ya Baroa.*

barwetsana */ba-rwe-tsa-na/ (lebitso)* **BAPISA banana**

basebedisi */ba-se-be-di-se/ (lebitso)* batho ba sebedisang dintho: *temoho ya nako e fetileng e bontshitse hore bahlahlobuwa ba hlokang boiphihlelo bo lekaneng jwaloka basebedisi ba puo ho ka etsahala hore ba se etse mosebetsi wa boemo bo hlokehang.*

basebeletsi */ba-se-be-le-tsi/ (lebitso)* batho ba etsetsang ba bang mosebetsi: *kopano yena e tlo tshwarwa le basebeletsi ba tsamaisang mananeo a tshehetso a dipetlele.*

basebetsi */ba-se-be-tsi/ (lebitso)* **BAPISA basebeletsi**

basebetsimmoho */ba-se-be-tsi/ (lebitso)* batho ba sebetsang mmoho sebakeng se le seng: *kgatiso ena e lekotswe ke basebetsimmoho kaofela mme ba dumellane ka yona.*

basetsana */ba-se-tsa-na/* **SINONIME banana**

bashemane */ba-she-ma-ne/ (lebitso)* bahlankana ba mophatong ba mothating wa ho bolotswa: *fihla nako, ha kena bashemane ha tswelane pelo hose ngwana motle,*

bashenyana */ba-she-nya-na/ (lebitso)* **SHEBA bahlankana**

Bashweshwe /ba-she-shwe/ *(lebitso)* theneketso ya setjhaba sa Basotho se neng se okametswe ke Morena Moshweshwe: *mphemphe e a lapisa, bashweshwe.*

Basia */ba-si-a/ (lebitso)* seboko sa batho ba anang katse: *lerumo la torobela Basiya, Batobatsi, lerumo la kgatsima mollo.*

Basothong */ba-so-tho-ng/ (lehlalosi)* moo Basotho ba bokaneng kapa ba ahileng ka bongata teng: *shebang mona, Basothong ha hona makwala.*

bata *(lehlalosi)* ho phola kapa ho ba le mohatsela o mongata: *ho nka nako e teletsana ha ho bata hore dikobo di futhumale.*

bathehi */ba-the-hi/ (lebitso)* batho ba thakgotseng ntho e itseng: *bathehi ba mekgatlo ba ka ipelaetsa ebang ditefo di sa finyelle pele ho nako e behilweng.*

batle */ba-tle/ (lekgethi)* tlhaloso ya botle ba tshobitsi ya batho: *ba batle barwetsana, dipalesa.*

bjwaramanya */bjwa-ra-ma-nya/ (leetsisamodumo)* ho dihela kgalase fatshe: *a nka dijana a di bjwaramanya majweng.*

boena */bo-e-na/ (lehlalosi)* ho ba le kamano ka tswalo: *a mamela mmae ha a mo lemohisa bohlokwa ba boena le phedisano.*

boetapele */bo-e-ta-pe-le/ (lebitso)*: noikarabelo ba ho tataiso: *ke ka boetapele bo tjhatsi ba hae re atlehileng ho etsa bonnete ba hore bohle ba etsa mosebetsi hantle, ntle le tieho ya letho.*

bofelo */bo-fe-lo/ (lebitso)* boemo ba ho qetela: *haeba o ka kgetha ho tlohela patlisiso pele re qeta, o ka kopa ho kgutlela ngakeng ho etsa diteko tsa bofelo.*

bofetoledi */bo-fe-to-le-di/ (lebitso)* mosebetsi wa ho fetolela ditaba puong e nngwe: *sepheo sa seboka ke ho ntshetsapele tsebo ya mosebetsi wa bofetoledi.*

bofifi */bo-fi-fi/ (lebitso)* nako ya masisapelo: *mabapi le koduwa e bakilweng ke toloko ya matshwephenene tshebeletsong ya bofifi ya presidente.*

bofokodi */bo-fo-ko-di/ (lebitso)* ho hloka matla le bohlale ba ho phethahatsa kapa ho utlwisisa dintho: *ditsebi di bona ho na le moo ho bileng le bofokdli kabong ya ditulo mme di eletsa hore ho etsollwe qeto e seng e entswe kgahlanong le diqeto tsa makhotla.*

bofuma */bo-fu-ma/ (lebitso)* maemo a tebileng a tlhoko le tlala: *ba kenya letsoho twantshong ya bofuma mme ho tla etswa meralo e lokelang tabeng ena.*

bohahlaudi */bo-ha-hlau-di/ (lesupi)* tlwaelo ya ho tsamaya dibaka tse fapaneng: *polokeho ya bahahlaudi mebileng ya naha ha e tsebisahale mme seo se ka se bewe mahetleng a mmuso.*

bokunyata */bo-ku-nya-ta/ (lebitso)* ketso eseng molaong: *selemo sena e ne e le sa ho qetela seo a tla amohele tjhelete ena ya bokunyata*.

bokwetlisi */bo-kwe-tli-si/ (lebitso)* mokgwa wa ho matlafatsa mmele ho itokisetsa papadi: *lekgotla le lemohile ho na le tlhoko ya bokwetlisi le boitokisetso ditekong le ditlhahlobong papadi ya bolo ya maoto.*

bola */bo-la/ (lehlalosi)* ho senyeha haholo hoo ntho ekekeng ya hlola e sebetsa: *ha o ka e siya motjhesong nako e telele, e tlo bola.*

bolatjha /*bo-la-tjha/ (lehlalosi)* ho hloka tshetiso ka hohlehohle, ha merero e tsamaya ho latela takatso: *ha a utlwa hore ba amohetse kopo ya hae, a qalella ho kgolwa hore tsohle di tla tsamaya bolatjha*.

bolatsi /*bo-la-tsi/ (lehlalosi)* ha sebetsa se leoditswe hantle: *ba ne ba tswhere disabole tse bolatsi.*

bolaya /*bo-la-ya/ (leetsi)* ho fedisa bophelo: *ke nahana hore o se ke wa robala mona kajeno, manong ane a itse a tla boela a tla, ba tla tloha ba tlo o bolaya.*

boledisa /*bo-le-di-sa/ (leetsi)* ho bua ntho kapa motho hantle: *ba boledisa matla a thapelo.*

bolehe /*bo-le-he/ (lebitso)* ho tshaba ntwa kapa ditaba: *bolehe ba hae bo mo nyatsisitse ka batho.*

bolemi */bo-le-mi/ (lebitso)* mosebetsi wa ho lema le ho iphedisa ka temo: *bolemi ba kajeno bo a fokola ka lebaka la kgaello ya pula.*

boleng */bo-le-ng/ (lebitso)* matla le molemo oo motho kapa ntho e o jarang: *kopanyetso ya boleng menehelo ya bona ha e kopantswe kaofela.*

bollotsa /*bo-llo-tsa/ (leetsi)* ho isa bahlankana thabeng balo sehwa letlalo la botona ba bona: *ha ngata ke mang ya nkang qeto ha banna ba dilemong tsa hao ba batla ho bollotsa?*

boloke */bo-lo-ko/ (lebitso)* mantle a diphoofolo: *ho se tlole ka meriana ya lapeng/setso (jwalo ka boloko kapa molora).*

bolotsana */bo-lo-tsa-na/ (lebitso)* ho se tshepahale: *e tlile le melawana e mengata mme emeng ya yona e tlisitse lesedi ha emeng yona e tlisitse pherekano le bonohe hamoo le bolotsana.*