The compilation of corpus-based Setswana dictionaries

Fannie Sebolela

A thesis submitted in accordance with the requirements in the field of study: DLitt: African Languages at The University of Pretoria.

December 2009.

Supervisor: Prof. D.J. Prinsloo

Summary

The aim of this thesis is to describe how corpus-based Setswana dictionaries should be compiled. The challenge to the modern Setswana lexicographer is to compile very practical descriptive and user-friendly dictionaries. A detailed evaluation of existing Setswana dictionaries will be performed in terms of the macrostructural and microstructural aspects:

- Coverage of frequently used words.
- Effective use of dictionary space.
- Use of standard dictionary conventions.
- Choice, ordering and composition of translation equivalent paradigms.

The focus will be on material collection and corpus building. Informants will be used to compile an oral corpus of 100,000 tokens. All ethical requirements such as informed consent requirements (See Appendix 1) will be honoured. Since the text corpus is an organic corpus, thus not a designed corpus aimed at balance and representativeness, the oral corpus will be constructed in the same way i.e. only basic selection criteria:

- Mother tongue speakers of Setswana.

- Adults (to be on a par with authors of the written sources in the text corpus). Age: ranging from 20-60 years.

- Male and female.

Critical analysis of all currently available Setswana dictionaries will be done with special reference to the dictionaries of Brown (1987) (SESD), Snyman, et al. (1990), Matumo (1993).(MSED), Kgasa (1976) (THAND) and Kgasa and Tsonope

(1995).(THAN) In all these cases the strategy would be in terms of the theoretical criteria and best practices in terms of a broad theoretical survey of core aspects of dictionary compilation. Finally, the study will be concluded with an analysis of corpus integrity and stability of Setswana corpora based on the model introduced by Prinsloo and De Schryver (2001a).

Declaration

I declare that **The compilation of corpus-based Setswana dictionaries** is my own work and all sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Fannie Sebolela

Acknowledgements

I am deeply indebted to Professor D.J. Prinsloo for his constructive criticism, purposeful guidance and patience without which this study would probably have been abandoned. May the Lord God preserve you.

My sincere gratitude also goes to Professor P.M. Sebate, Professor M. R. Malope, Dr R. Noormohamed, Ms. R. Ramagoshi, Ms. V. Nicolson and lastly, my dear sister Ms. C. C.L. Ngobeni for correcting and sharing some ideas in this research.

I am also grateful to Mr. M. Phaladi for his willingness to assist in the technical IT support.

Lastly, to my late wife Mmemme and our beloved children, Neo, Motumiseng, Omolemo and my grandson Motheo: You have been there for me from day one. This was a rough one, but finally I have completed the research.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1. Background to the study

It is a generally accepted fact that most dictionaries for African languages are not products of high lexicographic achievement. Gouws (1990:55) maintains that:

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

Snyman et al. (1990: preface) highlight another problematic aspect of dictionaries compiled without the use of a corpus namely that the dictionary team is aware that common and even essential words may easily be omitted during the compilation of a dictionary. The reason for this can be because the dictionary compiler had not encountered such words before, and merely hope that there are not too many examples of this kind.

For Setswana in particular, the dictionaries that are currently available are products of limited efforts with a minimum standard of lexicographical achievement. Typical examples are the inconsistencies found on the macrostructural level and inferior treatment of lemmas on the microstructural level. Many of these inconsistencies can be attributed to not making use of written as well as spoken electronic corpora that have become available to the lexicographer during the past two decades. If African lexicography is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should become an absolute priority, cf. De Schryver and Prinsloo (2000a and 2000b) and Prinsloo and De Schryver (1999 and 2001b).

Modern-day lexicography emphasises the importance of being guided by the user-perspective when compiling dictionaries. However, consideration of the user-perspective does not seem to play a role in most dictionaries of African languages.

A user-oriented dictionary should lead to enhanced information retrieval procedures. (See Prinsloo and Gouws 2000:138). The user-perspective, so prevalent in modern-day metalexicograghy, compels lexicographers to compile their dictionaries according to the needs and research skills of well-defined target user groups. It is believed that the dominant role of the user has had a definite effect on the compilation of dictionaries as well as on the evaluation of their quality. According to Prinsloo (1994:93), previously the user-perspective was not seriously considered in the compilation of African language dictionaries in general and Setswana dictionaries in particular. The role of the lexicographer was, to a large extent, seen to prescribe to the users what was presumed to be the ‘correct’ meaning and use of words. The challenge to the modern Setswana lexicographer is to compile practical, descriptive and user-friendly dictionaries. Good dictionaries can be used as linguistic instruments by their respective user groups. The best dictionaries are those that are the most user-friendly (see Gouws and Prinsloo 2005).

The mission of the lexicographer should be to enhance the quality of dictionaries along the way to the perfect dictionary.

According to Prinsloo and Gouws, (1996:103), the lexicographer is the mediator between linguistics and the dictionary user. Given the complicated morphosyntactic system of African languages, the task of the lexicographer as mediator is even more challenging.

1.1 Aims of the study

As so far no attempt was made to design a comprehensive, modern day approach for Setswana lexicography, this study aims at providing the necessary knowledge, in terms of implementing internationally accepted standards to the specification of African languages. Firstly, it is hence aimed at delivering a broad theoretical description of issues of today's dictionary compilation. Secondly, current African language dictionaries will have to be compared to this standard in order to demonstrate in detail where the African languages lexicography approach still has to be developed. Our general aim is to deliver a guideline for lexicographers on how to compile corpus-based Setswana dictionaries in combining both issues.

The aims of the study can be divided into three categories:

- a broad theoretical survey of core aspects of dictionary compilation.
- a critical evaluation of African language dictionaries with specific focus on Setswana dictionaries.
- an attempt to design a comprehensive lexicographic approach for the compilation of corpus-based Setswana dictionaries.

These aims will now be analysed in more detail:

1.2 Broad theoretical survey of core aspects of dictionary compilation

The aim is to describe the principles and practice that are characteristic of the compilation of modern dictionaries and then to evaluate dictionaries of the major languages of the world such as English, French and German, according to these principles. Issues like the collection of textual data, corpus compilation, access structure, and macrostructural and microstructural aspects will all be examined. Specific emphasis will be placed on the macrostructural aspects such as the compilation of the lemma list in the central text and the design of front matter and back matter texts, since these are regarded as the most common shortcomings in African language dictionaries (De Schryver and Prinsloo, 2000b and 2000c). On the level of the microstructure, the focus will be, inter alia, on clear and comprehensive sense distinction, and on the selection of appropriate examples. Lastly, we will examine definitions (paraphrases of meaning) of a high lexicographic standard.

1.3 Critical evaluation of African language dictionaries with specific focus on Setswana dictionaries

The principles and practices that are characteristic of the compilation of modern dictionaries and knowledge gained from the evaluation of the dictionaries of major languages of the world, as mentioned above will be applied in a critical evaluation of African language dictionaries with specific focus on Setswana dictionaries. This evaluation will be focused on the macrostructural and microstructural aspects as given in the background study of this chapter with special attention to the lexicographic aspects unique to African languages such as tonal indication and lemmatization strategies. Critical analysis of all currently available Setswana dictionaries will be done with special reference to the dictionaries of Brown (1987),(SESD) Snyman, et al. (1990),(SEAD) Matumo (1993),(MSED) Kgasa, (1976) (THAND)and Kgasa and Tsonope (1995).(THAN)

1.4 Designing a comprehensive lexicographic approach for the compilation of Setswana dictionaries

The research outcomes in 1 and 2 will finally be utilized with focus on the compilation of macrostructures and microstructures of high lexicographic standards for both monolingual and bilingual dictionaries of Setswana.

Conclusions reached from the study of core literature. (Zgusta 1971, Landau, 2001, Gouws and Prinsloo, 2005) and the preliminary analysis of Setswana dictionaries, indicate that the major shortcomings in Setswana dictionaries are:

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

The envisaged research aims are therefore extended to the compilation of corpus-based Setswana dictionaries that can be broken down into the following phases:

- build an oral corpus for Setswana.
- enlarge the current Pretoria Setswana text corpus.
- design lexicographic tools such as a ruler and block system for the oral as well as for the extended written corpus.

1.5 Research methodology

The study will:

- present a comprehensive theoretical conspectus of the creation of electronic corpora followed by a practical exploration for the African languages.

- focus on various applications of lexicographic principles and practices in the broad field of lexicography.
- use informants for the compilation of oral corpora, such as mother-tongue speakers of Setswana, as sources of spoken data, which will be processed to form an oral corpus.

1.6 Scope and limitation of the study

The study consists of six chapters.

Chapter 2 will focus on the origin and historical development of Setswana as a language. It will further highlight the missionaries's contribution and the role played by various stakeholders such as the Setswana Language Board towards the development of Setswana. The chapter also indicates how other languages or dialects have affected or influenced the orthography of Setswana and addresses the challenges the Language Board faced in the standardization of Setswana. Finally, the chapter illustrates how an oral Setswana corpus was compiled and how language changes led to new terminology being developed and eventually enriching the lexicon of the Setswana corpus.

Chapter 3 first presents a comprehensive theoretical conspectus of the creation of electronic corpora followed by a practical exploration of the African languages. It also focuses on various applications of lexicographic principles and practices in the broad field of lexicography and outlines the different steps needed to compile a structured corpus. This corpus should in due course, take an organic form containing texts from the genres and domains used including those that are spoken and written in private as well as in public. Furthermore the chapter highlights how corpora are mainly queried to obtain or generate word lists reflecting overall and comparative counts or concordance lines giving keywords in context. Prinsloo (2004:42) regards the keyness function in WordSmith Tools, for example, as ideal for selecting the so-called keywords by comparing, for instance, a dedicated Setswana corpus with a general corpus. Illustrative contrasts between the South African oral and text corpora and the Botswana text corpus will be demonstrated using the frequency word lists. Finally, studies will be conducted to monitor the stability of the growing organic Setswana corpora.

Chapter 4 will highlight gaps and inconsistencies in Setswana dictionaries on the macrostructural level. Specific emphasis will be placed on macrostructural aspects such as lemma lists in the central texts and frequency indication strategies in the dictionary, given that these aspects form the most common shortcomings in African language dictionaries, particularly in Setswana dictionaries. Frequency lists, culled from corpora, will be used to evaluate the lemma lists on the macrostructural level as well as for frequency indication.

Chapter 5 will focus on the importance of how the utilisation of a corpus can enhance the quality of the microstructure of the Setswana dictionaries' articles. Four major issues around the importance of corpora will be covered. Firstly, corpora as an aid to sense distinctions. Secondly, corpora as an aid to retrieve typical collocations. Thirdly, corpora as an aid to pinpoint frequent clusters. Finally, as an aid to select typical and natural examples. The chapter will also highlight problems related to the treatment of verbs, polysemous multiple meaning, synonyms and the treatment of the Setswana months in the currently available Setswana dictionaries. Thus, each section will conclude with suggestions for the improvement of the respective Setswana dictionaries by means of a corpus-based microstructure.

Chapter 2

Origin and development of the Setswana language

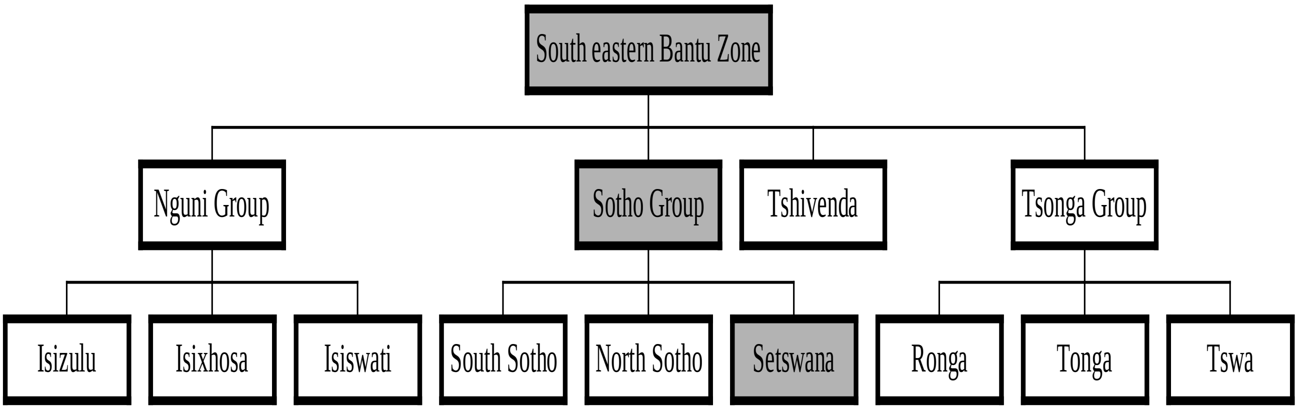
2.1 Introduction

This chapter deals with the origins and the historical development of the Setswana language. Various missionaries and other contributors to the development of the Setswana language will be discussed with emphasis on how their involvement led to the evolution of the language. As the Batswana are widely spread in southern Africa, the influence of various residential environments on their language will be highlighted with the aim of illustrating the emergence of the various Setswana dialects. Maps, illustrations and graphs, will be used to help interpret and comprehend the dynamics of this language. This study will also illustrate how oral Setswana came to be recorded and eventually compiled in grammar books, the Bible and/or hymn books and dictionaries.

No language is stagnant; all languages are affected by various environmental factors giving way to the development of a new dialect, which leads to vocabulary and grammatical changes in the original language. The growth and the development of Setswana necessitates that other language development processes be looked into to make a quantitative comparison between Setswana and other languages. The challenges faced in the development of Setswana will also be illustrated and special reference will be made to the role played and is yet to be played by various stakeholders e.g. the Language Board. The opportunities that are available to the stakeholders in South Africa will be highlighted.

2.2 Historical background

Setswana is one of the nine African Languages spoken in South Africa and is one of the Sotho sub-families of the Bantu language groups. Schapera (1953:14–15) maintains that there is no definite information about the origins of the Sotho peoples, to whom the Batswana belong. It is believed that they broke away from the main body of the Bantu-speaking peoples in and around the area of the Great Lakes in East Africa, and that they could have entered the region of today's Republic of South Africa through the western parts of southern Zimbabwe (previously southern Rhodesia) in three series of migrations. The first group were the Kgalagadi who settled in Botswana (previously Bechuanaland); the second group were the ancestors of the present Barolong and Batlhaping who first settled in the upper part of the Molopo river and later spread to the south and east; and the last and biggest group, which is regarded as 'the ancestors of all the Basotho tribes' settled with the other groups in the south-western part of Limpopo (then northern Transvaal), (http://Setswana.blogspot.com).



(http: Setswana. Blogspot. com com)

Figure 1: Graphical representation of the South-eastern Bantu Zone

It is a well-known sociolinguistic fact that an attachment to a certain language or dialect may be used to signal group solidarity. Given the current available data, we can only conclude that the compelling body of evidence points to the fact that

Setswana has, for at least two centuries, been the body label that includes the various Setswana dialects. According to missionaries, Setswana appeared to be by far the most extensively spoken language in southern Africa comprising a variety of dialects.

These united groups later broke into separate clusters, the biggest being the Bahurutshe, the Bakwena and the Bakgatla. What Schapera (ibid) further emphasizes is that the Batswana had already occupied the eastern half of their present tribal area by A.D. 1600. However, the tribes split constantly owing to constant infighting due to chieftainships and other factors. At present, the Batswana dialects are divided into four sub-groups (Cole, 1955: xvi-xix), each with their own various dialects. The sub-groups are:

(a) Central Setswana

The central Setswana includes the Serolong, Sehurutshe and Sengwaketse dialects.

(a) Serolong is divided politically into four sections:

- o Barolong boorraTshidi spoken in the Mafikeng district.
- o Barolong boorraRratlou spoken in Khunwana, Kraaipan and Setlagole.
- o Barolong booRapulana spoken in Lotlhakane and Polfontein (south-east of Mafikeng).
- o Barolong booSeleka with headquarters in Thabanchu in the Free State. (Even though the neighbouring Sesotho influences this dialect, it retains the characteristics of the central division.

- Sehurutshe spoken at Zeerust in the Marico district.
- Sengwaketse spoken at Kanye in Botswana.

(b) Southern Setswana

The southern Setswana sub-group includes the Setlhaping and Setlhware (Setlharo) dialects.

(a) Setlhaping is spoken in the Taung, Vryburg and Barkley West districts of the Cape Province.
(b) Setlhware is spoken in the Kuruman district.

(c) Eastern Setswana

Eastern Setswana comprises:

- Eastern Sekwena, spoken in Brits, Swartruggens, Rustenburg and Ventersdorp
- Sekgatla, spoken at Mochudi and Moshupa in Botswana, and in the Pilanesberg and Hammanskraal districts in Gauteng. The Sekgatla dialect is further divided into the following sub-dialects:
  - Sekgatla sa ga Kgafela
  - Sekgatla sa ga Moche
  - Sekgatla sa ga Mosetlha
  - Sekgatla sa ga Mmakau
  - Sekgatla sa ga Mmanaana

From the above-mentioned sub-dialects, Sekgatla sa ga Mosetlha and Mmakau are the sub-dialects spoken in and around Pretoria. The language with which they are in constant contact is predominantly Sepedi.

Tsonope (1990:33-35) concentrates mainly on the five major Setswana dialects found in Botswana. These are:

- Sekwena, used in the central eastern part of the Kweneng district (previously known as Botswana).

- Sengwato, used over the largely populated area known as the Central District.
- Sengwaketse, used south of the Kweneng, which makes up the Southern District and shows some influence from the Sekgalagadi (Khoi).
- Setawana used near Lake Ngami, not far from the Okavango swamps, known as the North Western District.
- Sekgatla, used in the area called the Kgatleng District.

Tsonope states that there are other numerically far smaller groups such as the Batlokwa, located to the east and south of Botswana, who were originally the Bapedi, speaking the Sepedi language. The Barolong live in the southern district of Botswana and the Batlhaping live further south in the district of Botswana.

The Setswana dialects spoken in South Africa show the influence of English and Afrikaans (the previous official languages of South Africa that were imposed on Batswana students before the change of government in 1994, this includes the use of foreign words) non-changed, as there is no expression in Setswana for these terms and the implementation of words into Setswana as loan words. Currently, Setswana is one of the eleven official languages in South Africa.

In urbanized areas, Setswana is also strongly influenced by the neighbouring languages but in areas where Setswana is predominantly spoken, there is little influence from other languages. Languages in the north-west areas are spoken by small groups of other African tribes such as Sepedi, Sesotho, IsiZulu, IsiXhosa, SiSwati, Tshivenda and foreigners from neighbouring African countries, like Zimbabwe, who are employed in these areas. These smaller groups tend to speak their languages among themselves but, in public, they are forced to communicate in Setswana, Afrikaans or English.

Since research for this thesis was conducted in South Africa, the classification of Cole (1955), interpreting language facts from a South African perspective, is used as a basis in the discussion of dialects and the standardization of Setswana. The Setswana dialects can be constructed as follows:

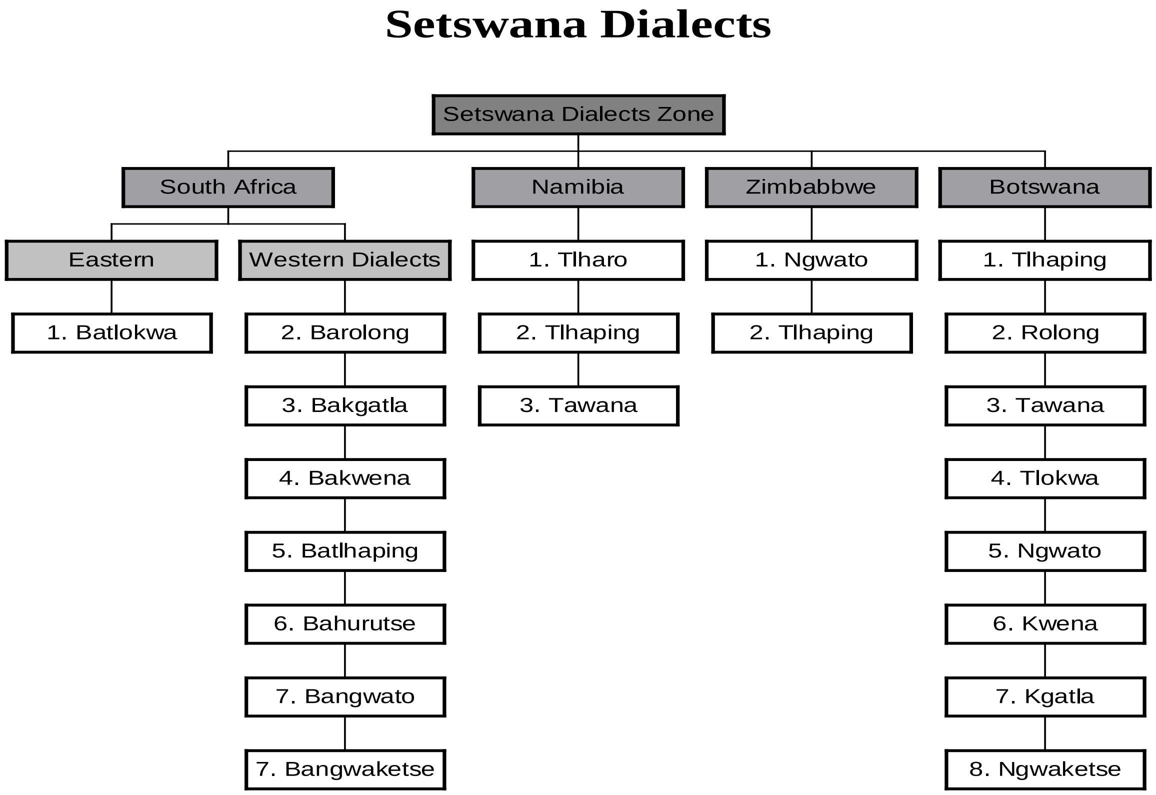
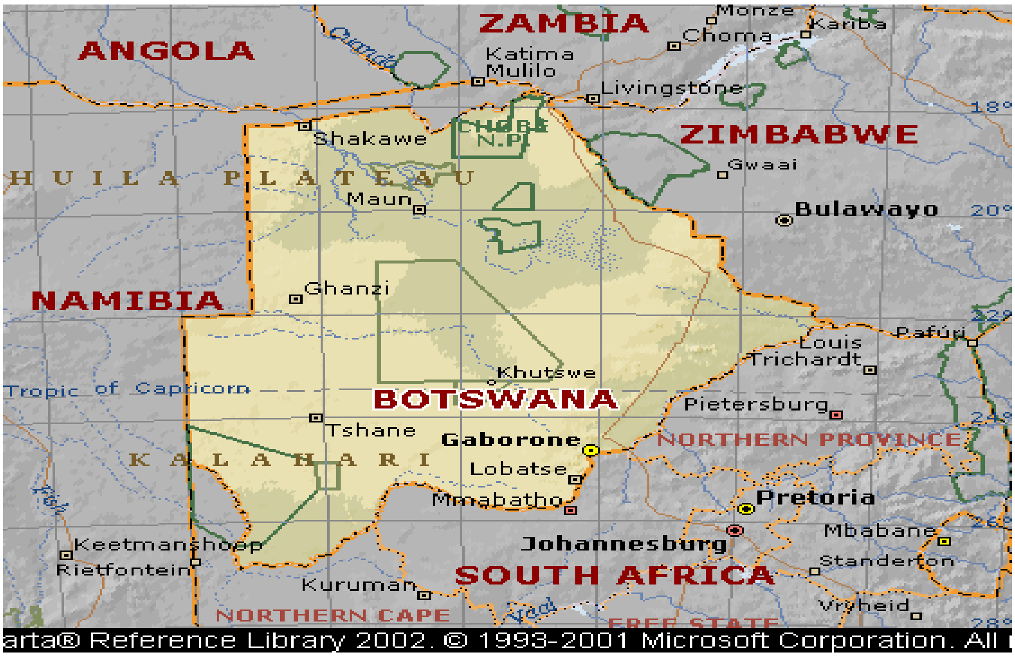


Figure 2: Setswana dialects

2.3 The role of missionaries towards documentation and the development of the language

According to Breutz (1989:10), missionaries in the 1820's who spoke Setswana, distinguished two sister languages namely, *Sichuana* and *Kaffir*. In their opinion, Setswana was more widely spread than the Nguni-languages along the Congo River because of the similarities in the grammar and basic vocabulary (Breutz, 1989:10). Breutz (ibid.) further highlights that Setswana appears to be the dominant language spoken in South Africa.

Setswana is the language most closely related to the Sotho language group, i.e. Setswana, Sesotho (Southern Sotho) and Sesotho sa Leboa (Northern Sotho). The Sotho-Setswana language group is bound in language and custom, and they happen to claim a common ancestor, Mogale. They also share certain family traditions, political and magical beliefs, and organizational culture. According to Breutz (1989:3), while the three Sotho languages are inherently intelligible for various political and historical reasons, they are generally considered to be three different languages.

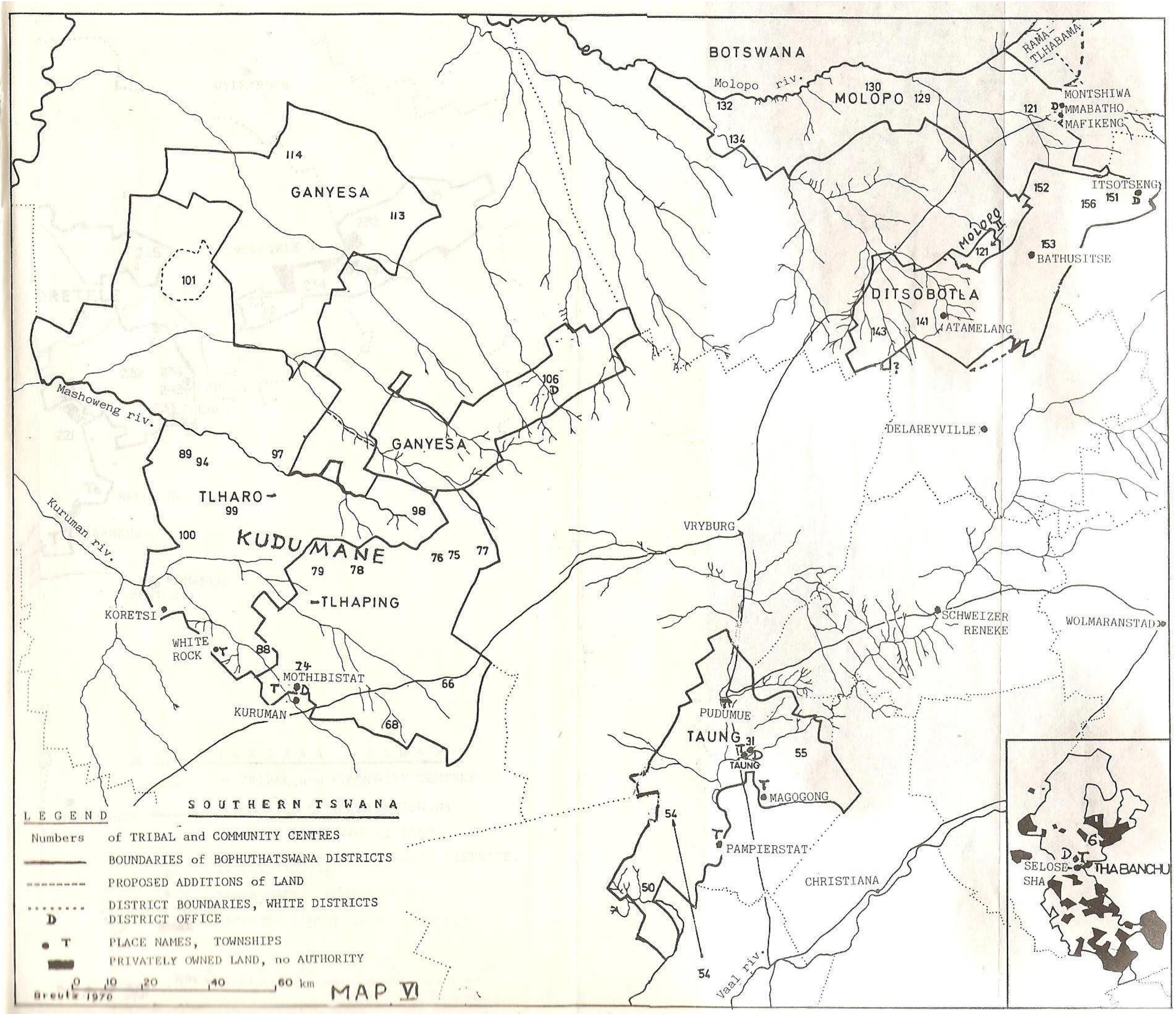


http://www.african.gu.sc/research/setswana.html,

Figure 3: Development of Setswana as a national language

Towards the end of the 17th century, the Setswana population increased greatly. Severe drought and famine during this period caused them to search for grazing and water. The Kwena tribe settled west of Rustenburg (now called Madikwe), and others (presently known as the Bangwato tribe) settled in Botswana, in Lobatse. The Barolong tribe migrated south to Taung and settled in Ganyesa. The Batlhaping lived on the banks of the Orange River where they lived on fish (*tlhapi* in Setswana), from where they received their name. They had contact with the Korana-Hottentots, and ended up settling in Vryburg/Kuruman.

Breutz, (1989:11) draws attention to the fact that the Batlharo migrated south of Ganyesa, and are presently also settled in Kuruman; the Bakgatla migrated to Hammanskraal, north of Pretoria (now Tshwane); the Bakgatla-ba-ga–Kgafela settled in the east; while the Batlokwa from the west came into contact with the amaNdebele tribe in Mamelodi and Eersterus, east of Pretoria. According to Ellenberg (Lit: 1912), a section of the Batswana tribe migrated to Lesotho for further grazing and called themselves the South Sotho.



(Breutz 1970: Map vi)

Figure 4: Distribution of the Southern Batswana tribes

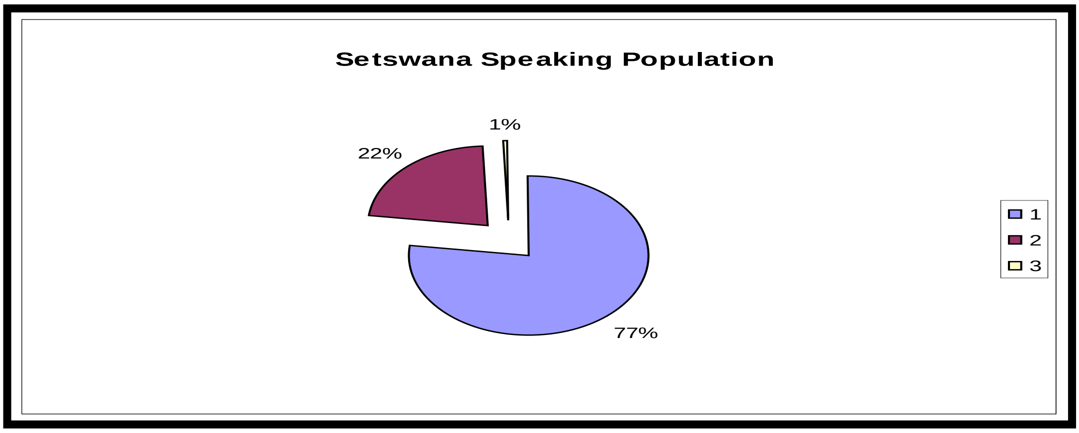
According to Breutz (1989), Rev. John Campbell, an English congregational minister, was appointed director of the London Missionary Society in 1816. His first work as a missionary was among the Batlhaping tribe in Ganyesa (formerly known as Vryburg). Rev. Robert Moffat later joined him and they began building a new mission station in 1824. The people with whom Moffat laboured at Kuruman were called the Batlhaping. They were the southern-most tribe of the Bantu-speaking group, collectively known as Batswana (also written *Bechuana* or *Bachmann*), and were the first group to encounter white people in the area of today's Botswana. Campbell and Moffat translated the Bible into Setswana and printed it at the mission station. From 1827 onwards, Methodist missionaries joined the London Missionary Society.

Cole (1992) and Ramagoshi (2000) are of the opinion that the Batswana was the first language group to receive formal schooling, except for the Xhosa in the south. Pupils of all ages were taught together in one class. In 1870, the first teachers were trained at the Bethanie mission station. Students were taught lessons from the Bible, catechism, singing, reading, arithmetic, geography, writing, Dutch and the method of teaching. Schools were built all over the country.

2.4 Population census of the Batswana speaking people

Setswana is one of the eleven official languages of South Africa, and is the national language of Botswana. According to the Report No: 03-01-11 (1996), there are approximately **3,677,016 (8,2%)** people in **South Africa**, an estimated **1,070,000** people in **Botswana,** and about **30,000** people in **Namibia** whose first language is Setswana.

The Batswana population can be illustrated as follows:



http://www.cyberserv.co.za/users/jang/unesco/setswana.htm

Figure 5: Census in brief: Report No: 03-01-11 (1996)

Key*:* Numbering on the graph.

1	=	South Africa.
2	=	Botswana.
3	=	Namibia.

2.5 Contribution by the missionaries

In summary, the missionaries pioneered the study of the language, devising orthography and systematizing their findings into a first version of a Setswana grammar. (Cole, 1992: xxiii). They defined Setswana as a disjunctively written language, i.e. a language where bound and free morphemes are usually written separately. Thanks to the extensive and detailed work done by the missionaries, in what we speak of today as the history of Setswana orthography, the Setswana language became the first written Sotho language in 1806, (Breutz, 1989:3–12).

Missionaries developed Setswana orthography rather according to their understanding, than according to the phonological rules given below. However, their contributions to the development of Setswana orthography are still significant. Ramagoshi (2000:9) describes the background for the following significant developments:

- In 1806, the German, Hinrech Litchtenstein, wrote 'The language of the *Beetjuana'*, which was later translated into English. He considered the Batswana to be a single linguistic group and wrote '*Beetjuana'* words. He used the prefix *Bee***–** instead of ***Ba* –,** which is the orthography of the prefix used today and the sound ***tj*** in contrast to the modern ***tsw,*** e.g. ***Beetjuana*** instead of **Botswana**.

- During his travels around South Africa in 1815, John Campbell devised a list of 80 'Bootchuana' words. He wrote '*Bootchuana'*, the prefix Boo- instead of ***Ba*** (prefix*)* ***tchua*** (root) – instead of (*tswa*). Burchell, who wrote about the Batswana in 1824, adopted this writing. Dr Robert Moffat from the London Missionary Society arrived among the Batlhaping in Kudumane in 1825. He used '***c****'* instead of ***tsh*** as in ***sh,*** for example, in writing *Moshe* in the Bible. Professor Jones and S.T. Plaatjie wrote books called the *Sechuana Reader* and the *Jones Secwana nouns* respectively, in 1929. In 1876, Rev J T Brown wrote the *Lokwalo loa Mahuku a Secoana le Seengelese*. According to Cole (1992) and Ramagoshi (2000) it should however be noted that Brown and Moffat were influenced by the Setlhaping dialect in their writing, while Archbell and Casalis wrote in the Serolong dialect. In the word *kgomo* 'cow'**,** Rev Moffat used ***kh*** in contrast to ***kg*** in today's orthography. He used '*banona*' in the place of banna. (UNESCO: 2000). Studies of the Batswana history traces the historical development of the major studies and function as a general reference work to the contemporary dictionaries.

2.6 Relationship between Setswana dialects and the standard language

In a descriptive synchronic sense, language can refer either to a single linguistic norm or a group of related norms. A dialect is then any one of related norms comprised under the general name "language", which is, historically, the result of either divergence or convergence. According to Haugen (1966:923), the two concepts are cyclically applicable, with language always the super ordinate and “dialect” the subordinate. For example, every dialect is a language, but not every language is a dialect.

In Setswana a dialect suggests neither informal, lower class nor rural speech, but is part of the language. According to (Haugen, 1966:926), there are two distinct dimensions involved in the various usages of language and dialect. One of these is structural i.e. descriptive of the language itself, the other is functional i.e. descriptive of the language its society uses in communication. Setswana has eight dialects; this means that there are eight identifiably different speech forms that have enough cognates to make it certain that they all developed from mutually understandable language. A language can also be functionally defined as a superposed norm used by speakers whose first and ordinary language may be different. Thus a language is the medium of communication between speakers of different dialects. Sociologists on the other hand may be referring to the fact that the language is more prestigious than the dialect while the dialect may be defined as an undeveloped language i.e. not standardized. South Africa is considered to be a country that uses more than one language for official purposes, but English is used as the international language (lingua franca).

In African languages, such as Setswana, the process of standardization was intimately tied to technological and scientific changes. These constitute a challenge for the language to achieve full development. The standard is threatened by the existence of rival norms, for example dialects among its users.

It is important to note that every vernacular, Setswana in particular, can at the very least add words borrowed from other languages, but usually possesses devices for making new words from its own resources as well. It is also important to note that where transitions in Setswana are gradual, it may be possible to find a central dialect that mediates between extremes, one that will be the easiest to learn and most conducive to group coherence.

The relationship between the standard language and its dialects described above has very specific implications for dictionary compilation for Setswana. Given the lack of a sufficient number of Setswana dictionaries aimed at specific target users, not to mention dictionaries for the different dialects or for productive or receptive use, the Setswana lexicographer is forced to accommodate the standard language as well as frequently used dialectical forms in the dictionary. There is no opportunity at this point in time to compile separate dictionaries aimed at individual dialects.

2.7 Language planning

Christopher Brumfit (1992:580) describes language planning as "the attempt to control the use status, and structure of a language through a language policy developed by a government or other authority". The Random House Dictionary of the English Language concurs, but adds some significant detail. Language planning is "the development of policies or programmes designed to direct or change language use, as through the establishment of an official language, the standardization or modernization of a language, or the development or alteration of a writing system". (See page 27-34).

Commentators usually distinguish between 'natural' and 'interventionist' language planning. Natural language planning actively supports the evolving language needs of a society as they emerge in response to other-than-linguistic pressures. It supports no major effort to encourage language shift or change which runs counter to the emerging language dispensation induced by other-than-linguistic changes in a society. Interventionist language planning, in contrast, is prepared to challenge the impact on the language dispensation of current sociolinguistic forces. It sets itself to revitalize moribund languages, preserves dialects, maintain languages that are under threat, modernize traditional languages for the use in different domains and defend language rights.

The term 'language planning' has often been identified within a third-world context as a way of standardizing national languages as a part of modernization and nation

building. It is therefore to be noted that language planning is indeed neither a modern phenomenon nor is it confined to the third world.
(http://en.wikipedia.org/wiki/language planning).

Cooper (1989) divides language planning into three sub-dimensions, as indicated below:

- **Corpus planning:** Tauli (1968) defines corpus planning as interventions in the forms of a language. Developing orthography and resources of a language to be used as a medium of communication for modern topics equipped with terminology to be used in administration, education, conversation comes under the sub-dimension of corpus planning. Exclusion of foreign words and spelling also come under this dimension. Therefore, all aspects of grammar, writing systems and standardization are concerns of corpus planning.

- **Status planning** refers to the deliberate efforts to allocate the functions of languages and literacy within a speech community. Once a suitable dialect has been chosen it will be elevated into an official language. Status planning creates a new writing system and locates the language and literacy within the community even at the expense of other competing dialects. It is mentioned that status planning is the most controversial aspect of language planning.

- **Acquisition planning** concerns the teaching and learning of languages, whether national, second and/or foreign languages. It involves efforts to increase the number of users and the distribution of languages by creating opportunities or incentives to learn the language, provincial or national, second or foreign language.

It is also important to note that planning is not necessarily conducted on a national level. Ethnic, religious or occupational groups can also carry it out. For example, by using new and emerging vocabulary to drive language change through our interactions, we pick up new words and integrate them into our daily speech.

2.8 Language change

Diamond (1993) states that in most cases, language change is gradual and occurs through the long course of the intergration of the society with another via political unification, mobility, intermarriage, or education. While societies and their languages continually change in response to internal and external circumstances, there are proactive measures that can be implemented to either maintain the particular direction in which that language is moving or reverse it.

Dorian contends that language change 'is the gradual displacement of one language by another in the lives of community members'. (cited in Huebner 1987). Huebner (1987) adds that language change can be either partial or complete. The following are some conditions that affect societal language change:

- Origin of the contact situation

  Voluntary migration, especially of individuals and families, results in the most rapid language change.

- Status differential in power and economics

  The official languages occupy a particular domain, namely that of administrations, government, education and the media, whereas indigenous languages are found primarily in the domains of family and friendship. Mansour (1993) claims that indigenous languages are more susceptible to language change because of their limited function of serving social interactional needs.

- Cultural Values

  Kulik (1994, p. 7) posits that ethnic identity or the way in which the expression of positive and highly valued aspects of the self comes to be bound to expression through a particular language is the most crucial factor influencing the rate and finality of language change.

- Demographics

  The size of the speech community reflects the vitality and potentiality for language change or maintaince. According to Crawford, (1995), movements of people will affect language use. Mobility often leads to intermarriage with other language speakers.

- Status of the Writing System

  According to Spencer, (1992), a writing system legitimates literacy efforts which in turn contribute to the cultural production and vitality of a community. There was a massive effort in Setswana to standardize the writing and spelling system and to develop dictionaries and grammars. (See page 28-35).

The language change is caused by various factors, for instance, sociolinguistic reasons, e.g. the imitation of English and Afrikaans as spoken by the upper classes, the historical dominance of correspondence in English, and so forth. Before 1994, the political boundaries set limits to the degree of intercommunication among people living in the same area, as was the case in the former Bophuthatswana, one of the so-called Bantustans, now the North West Province. The Bophuthatswana government restricted the use of English as a form of correspondence and introduced Setswana as learning and teaching medium from Grade 0 to Grade 6. The implementation of language development activities and literacy was developed and training in the translation and development of literacy materials in Setswana was encouraged.

According to Chaucer (1902), before a language can change, speakers must adapt new words, sentence structures and sounds, spread them throughout the community and transmit them to the next generation.

Lyons (1981:208) refers to two general factors of language change as ‘analogy’ and ‘borrowing’

(a) Analogy

> "An analogy can be defined as a spoken or textual comparison between two words to highlight some form of semantic similarity between them. Linguistics defines analogy as a process that reduces word forms perceived as irregular by remarking them in the shape of more common forms that are governed by rules. For example, the English verb ‘help’ once had the preterits *holp* and the past participle *holpen*. These obsolete forms have been discarded and replaced by ‘helped’ by the power of analogy."

Langacker (1987:445–447) states, ‘analogy’ is often presented as an alternative mechanism to generative rules for explaining productive formation of structures such as words. Others argue that they are, in fact, the same mechanism, that rules are analogies that have become entrenched as standard parts of the linguistic system.

Lyons (1992:30) defines ‘analogy’ as the main factor which the neo-grammarian invoked to account for exceptions to their sound-law[1] He states:

> "It came to be realized that ‘analogy’ was a major factor in the development of languages at all periods and could be attributed merely to periods of decline and corruption."

[1] The sound-law meant to aim at an orthography that reflects the phonetics of the language

It is important to note that the development of language had frequently been influenced by the tendency to create new word forms with more common and regular patterns of formation. In Setswana, 'analogy' can be defined as 'equivalents' in a language, which are determined by the frequent use of a particular word in a given environment. An example of high frequency is *go kgwa* and *go tlhatsa* (to vomit). The word *go tlhatsa* is more frequently used than *go kgwa*. The rule is also more 'regional' as the words go *kgwa* is more often used in the Free State as a dialect than *go tlhatsa*.

In many languages there are several stages of politeness when people interact. As such, a mother tongue speaker puts several of them into practice in everyday communication. This fact entails that words referring to the same entity appear in different forms in several internal lexicons of the speaker. Each of these forms is only to be uttered in a certain social interaction, dependent on the social roles of speaker and hearer and the context of their interaction, and the position they take in the social hierarchy. For example, some words may be explicitly used by women towards other women; others might be acceptable between teenagers but a taboo towards elders, like, e.g. *lesipa*, as in the following:

'***Lesipa****' la ntswa le dujwa le sa le metsi (*mess of a dog is stirred whilst wet)
'***mantle****' a ntswa a dujwa a sa le metsi (*mess of a dog is stirred whilst wet)

English meaning (Spare the rod and spoil the child)

'*Lesipa*' (mess) is an offensive, vulgar swearword and must not be used when speaking to adults or elderly people. It is argued that words such as '*lesipa'* encourage the youth to disrespect the values and norms of their community. The word '*mantle*' (mess) is more preferred culturally than *'lesipa*'. An analogy and its explanatory force in the case of exceptions to the 'sound- laws' can be illustrated as follows:

Ke 'tlaa' tla ka moso (I will come tomorrow)
Ke 'tla' tla ka moso (I will come tomorrow)

According to Setswana spelling rules, published in 2008, double vowels appear when '*a*' vowel occurs twice, as in future tense marker, *tlaa*: Previously, this future tense marker was spelt with one vowel as in *tla* (come) as in: *Ke tla tla ka moso*. (I will come tomorrow) and k*e tlaa tla ka moso*. (I will come tomorrow).

When looking at the behaviour of the '*tla*' and '*tlaa*' the sound '*tlaa*' seems to be more appropriate than '*tla*'. In this instance, the behaviour indicates whether it has been modified or not.

Maipolelo (telling about yourself) > Curriculum vitae (CV)
Maikao > Curriculum vitae (CV)

The word *maipolelo* (CV) was replaced by *maikao* by the Language Board since the word *maikao* was considered more relevant than *maipolelo.* In this section, the research is of great significance in advocating the inclusion or exclusion of analogy in the Setswana dictionaries. In the National Setswana lexicography Unit (NLU), it is still a huge debate whether to include the new words (standardized) or those used prehistorically.

(b) Loanwords / Borrowing words

Kruger (1965:6) defines loanwords as words that are not indigenous in a certain language, but that are adapted to the morphosyntactic rules of the language which are used by the speakers of that language. To meet the needs of dictionary users, it is essential to standardize and improve the Setswana corpus to express modern concepts. Adapted words like the following were introduced to provide the terminology needed in subjects such as maths, science and business.

Tafel	*(Afrikaans)*	→	*tafola*
Stoel	*(Afrikaans)*	→	*setulo*
Cent	(English)	→	*sente*
Cement	(English)	→	*semente*
Train	(English)	→	*terene*
Motorcar	(English)	→	*motorokara*
University	(English)	→	*univesity*

(Terminology and Orthography no 4 (1998)

The introduction of modern technology, language change and loanwords from other languages such as English and Afrikaans causes the emergence of new words for Setswana. Consider the following examples:

Cell phone	→	*Mogala wa letheka*	>	Translation into new term
Programme director	→	*Motshwara-marapo*	>	Translation into new term
Teacher	→	*Moruta-bana*	>	Translation into new term
Aeroplane	→	*Sefofane*	>	Extension of meaning of a previously existing word.
Computer	→	*Komputara*	>	Adaptation to sound-law, loan word.
Celsius	→	Celsius	>	No adaptation, foreign word.

Crystal (1985:183) defines a loanword as a linguistic unit, which has come to be used in a language other than the one in which it originated. Loanwords should be directly translated to further develop the language. Such activity facilitates the preparation of the educational materials, and, generally strengthens the Setswana language. The speakers are not aware of the fact that these words are foreign, but have become part of that language. For example; *mogala wa letheka* (cell phone) is usually used by illiterate people and seldom in informal gatherings and it does not represent the language as it is spoken by the Batswana people.

It is important to note that language planning does not take place in a vacuum. A variety of social factors should be taken into consideration. For example, loanwords are well understood and preferred by the young urban population more than by people living in the rural areas. The use of borrowed/foreign language depends on social factors such as strong cultural contacts, age, geographical area and education. Gumperz (1982:66) defines borrowing words or loanwords as follows:

> "…the introduction of single word or short, frozen idiomatic phrases from one variety [i.e. language] into the other. The items in question are incorporated into the grammatical system of the borrowing language”.

According to Ramagoshi (2000:38), the Setswana Language Board constituted in 1960 and 1980 by the Department of Education, was the main body focusing on language development and standardization. For example, the Board focused on the translation of government documents including forms, notices and departmental reports.

The Setswana language cannot afford to ignore or neglect foreign influence. In this regard language planners often prefer borrowing or loaning from foreign sources. Thus modern dictionaries should play an important role in the implementation action of new terminologies.

2.9 Language development

Leepiles (2003) defines the purpose of language development as a domain that develops the Setswana language for use in scientific, technological and socio-economic fields and to inform the Batswana in the national language at government level and to promote the use of indigenous language systems. This includes the development of (1) a writing system (2) a dictionary and (3) grammatical rules. It is important to note that (1), (2) and (3) urgently need to be updated in dictionaries for Setswana.

2.9.1 Writing system orthography (1929–1937)

The central committee was disbanded and, in 1930, the Transvaal Sotho committees were formed with a Setswana sub-committee known as 'The Practical Orthography Committee'. During the same year representatives from the four provinces: Botswana, Free State, Cape Province and Transvaal (now Gauteng) agreed on using a writing system known as the 'Mafikeng Orthography'. In 1931, they devised another orthography, which differed from both the Mafikeng and the Central Orthography. During the 1937 conference, the differences and similarities between the Central Orthography and the Bloemfontein Orthography were discussed and the following was agreed upon:

- The semi-vowel would be written as ***y*** and ***w***
- The aspiration of affricatives would be written as ***h***
- The velar fricative would be written as ***g*** and no longer as ***x***

In the same year (1937), the Setswana Language Board was formed by the Transvaal Education Department. The two Sotho language boards, i.e. the 'Central Orthography' and 'Bloemfontein Orthography', were combined and the following proposals were made:

- Diacritics would be utilized for technical and scientific work only
- The vowels ***ê*** and ***ô*** would be used without diacritics as ***e*** and ***o***
- The ***g*** would represent ***x; g*** and the **h** would be retained for South Sotho
- Semi-vowels would be written as ***y*** and ***w***

(Bantu Studies, 1937:11-137:148. A practical Orthography for Tswana)

2.9.1.1 Articulator phonetics

Phonological information, such as vowels and consonants, are not provided in most Setswana dictionaries. Thus the forthcoming corpus-based dictionaries should attempt to include them. Consider the following:

(a) Consonants

The following articulations are placed among those used in Setswana and other related languages: (with or without secondary articulations of various kinds). We use the International Phonetic Association (IPA) alphabet for illustration.

- **Bilabial,** the lips are brought together; for example: **[p]; [b]; [m]**
- **Labiodentals,** the lower lip is brought into contact with the upper teeth, for example: **[f], [v]** whereas **[p]** , **[b]** ,[**m]** are stops**, [f]** and **[v]** are fricatives.
- **Dental,** the tip of the tongue is brought into contact with the upper teeth, for example: **[t] , [d] , [n] , [l] , [r]**
- **Alveolar,** the tip of the tongue is brought into contact with the alveolar ridge e.g. **[t] , [d] , [n] , [s] , [z]**

- **Velar,** the back of the tongue is brought into contact with the velum or soft palate, e.g. **[k]** , **[g]** and the **fricatives [x] and [ɣ]**
- **Glottal,** the vocal cords are brought together e.g. **[h]** and **[Ћ]**

(Terminology and Orthography no. 3, 1972 and Cole, 1992:19)

The Sotho Language Board of the former Transvaal Department of Education made proposals to change certain consonants which were more related to the Sotho language than the Setswana language. The following consonant changes were made to the Setswana Orthography:

Comparative list orthography

1929	**1950**
by	***bj***
fs	***fš***
fy	***fš (f , šw)***
x	***g***
hl	***hl (tlh)***
kx	***kg***
py	***pš (pš)***
phs	***phš***
psh/phy	***pšh***
ths	***tšh***

(Terminology and Orthography no 4 1988)

In the light of the above changes and the impact they will have on the effectiveness of a dictionary, it is recommended that the existing Setswana dictionaries be revised within the limits of their intended function.

(b) Vowels

a: *rata* (like)

***ê*: *r* *ê*ma**	*(chop)*	phonetically ***ê***: *r**ê**ma*
***e*:** *l**e**ma*	*(plough)*	phonetically **[e]**: *thabeng*
***į*:** *r**i**pa*	(cut)	
***ô*:** *l**ô**ra*	(dream)	
***o*:** *n**o**ka*	(river)	
***u*:** *r**u**ta*	(teach)	

Here are two points to bear in mind:

- The letters **e** and **ê** each represents two different sounds namely **[*e*]** and **[ɛ], [*o*]** and **[ô]** that are variants of each sound.
- The lengthened vowel (which normally occurs in the penultimate syllable) is written as a single letter: *pol**e**lo*; *ba rata n**a**ma; re **a** ja*

(c) Semi vowels

(a) Nasals ***m*** (lo***m***a); n (***n***ama), ***ny*** (n**y**ala), ***ng*** (***ng***apa)

There are two significant aspects to note on these nasals:

- Phonetic symbols are not to be used for the writing of ***ny*** and ***ng***, just because people have become used to ***ny*** and ***ng*** which function as syllabic nasals and are both represented by ***n***, as in ***n***nyala and ***n***ngapa

(d) Bilabial consonants

p (***p***ula), ***ph*** (***ph***ula), ***py*** (bo***pya***), ***phs*** (***ph***sina), ***phš*** (***phš***atla*)*, ***phy*** (***ph*y**a***ph***ya).

Note the following:

- **ps**, **py**, **phs**, **phy** and **phš** are called plosive consonants, and not fricatives.

- The committee points out that ***phs*** is a variant of ***phy.***
- Aspiration is indicated by the letter ***h*** which appears between the letters ***t*** and *s/š/y* [***tapa]*** ***[tanya]***

(e) Alveolar consonants.

t (tau), ***th*** (***th***aro), ***ts*** (***ts***ebe), ***ths*** (thsebe), ***tl*** (***tl***oga)
th, ***ts***, and ***thš*** are called **plosives**, and not **fricatives**.

(f) Velar consonants

k (bo***k***a*), kh* (***kh***umo), ***kx*** (***kg***omo)

(g) Fricative consonants (voiceless)

(i) Bilabial consonants

f (*fofa),* fs (l*efsifsi*), fy (*fyega*)

(ii) Alveolar and palatal consonants

s (*setlhare*), ***s*** (***s***upa*),* ***s*** (**s**ol

(Terminology and Orthography No. 4, 1988)

2.9.2 Grammatical rules

According to Lyons (1981:109), part of speech: nouns, verbs, adjectives, prepositions, abbreviations, etc. play a crucial role in the formulation of the grammatical rules of a language. The Setswana language is not an exception. Only abbreviation rules as an example will be discussed in this section. It is important to note that abbreviation plays an important role as follows:

- ❖ Abbreviation is a welcomed device especially in those varieties of language where time or space is at a premium.
- ❖ Abbreviation makes listeners and readers work harder, each element we omit, and each item we introduce, imposes an extra demand on listeners and readers.

2.9.2.1 Abbreviation rules

Current available Setswana dictionaries do not include the abbreviation rules. They should be indicated in forthcoming Setswana dictionaries. According to Lyons (1981:20), the generativist sees language change in terms of the addition, loss or recording of the rules that determine a speaker's linguistic competence.

There are a number of rules that could be applied:

- By using the first letter of each content word forming the description. For example:

LBPS	***L**ekgotla la **B**osetšhaba la **P**uo ya* **S**etswana
BABD	***B**iro ya **A**forika-**B**orwa ya **D**ikelotheo*
MPAB	***M**ophato wa **P**hemelo wa **A**forika-**B**orwa*
DSAB	***D**itirelo tsa **S**epodisi tsa **A**forika-**B**orwa* (South African Police Services)

- By using a small letter with a full stop after every letter. For example:

k.g.k.	*ka gore ke*	(for that reason)
k.j.	ka jalo	(therefore)
k.m.a.m.	*ka mafoko a mangwe*	(in other words)
j.l.j.	*jalo le jalo*	(etc.)

- In case of cardinal points, only the first and third letters are taken as abbreviations. For example:

Bk	*Bokone*	(North)
Br	*Borwa*	(South)
Bp	*Bophirima*	(West)
Bt	*Botlhaba*	(East)

- **Measurements.** It was agreed at the conference that the international abbreviations used in English should apply. For example:

l	*litara*
kg	*kilograma*
g	*gerama*
kg	*kilogerama*
m	*mmitara*
cm	*sentimetara*
ha	*heketara*

- **Acronyms**: Acronyms that are familiar and associated with people's lives cannot be changed. For example:

HIV	Human Immunodeficiency Virus
Aids	Acquired Immune Deficiency Syndrome
TB	Tuberculosis
JSE	Johannesburg Stock Exchange

- **University qualifications**:

SEC	Secondary Education Certificate
JC	Junior Certificate
BA	Bachelor of Arts
BSc	Bachelor of Science

LLB	Bachelor of Law
B.Com	Bachelor of Commerce

- **Personal names**:

Moprf	*Moporofeta*	(prophet)
Mot	*Motlotlegi*	(honourable)
Mokan	*Mokanseliri*	(counseler)
Mopr	*Mopresidente*	(president)
Mohn	*Mohumagatsana*	(mrs)
Ng	*Ngaka*	(doctor)
Mor	*Morutabana*	(teacher)

(Terminology and Orthography no. 4)

If the lexicographers are concerned about the user's needs, they simply have to consider improving the functional quality of the dictionary. The above-mentioned rules and vocabulary as applied in Setswana orthography require an urgent change in the macrostructure of the Setswana dictionary.

2.10 Somerset House Conference of 1947

Aims and Objectives

- To remove the differences between the Sepedi and Setswana orthography

- To use diacritics only in cases where they differentiate meaning and in technical and scientific publications
- To standardize the Sepedi and Setswana orthography

Committee proposals

- Diacritics should only be used in technical and scientific work
- The vowels e and o should be written without diacritics as in ***ê*** and ***ô***
- The letter g should represent the letter ***x***, g, and ***h.*** The letter *h* would be retained for South Sotho
- South Sotho should use ***d*** and not ***l*** as in Sepedi and Setswana.
- The letter ***g*** should represent ***x***, ***g*** and ***h.***
- The letter ***s*** should represent the ***s*** sound and, the letters ***[sh]*** should represent ***[š]***

(Transvaal Sotho Orthography of 1929 and the Setswana Orthography)

2.10.1 Impact of Bantu education on Setswana orthography

In 1962, the Setswana Language Board was formed and its duties included coining new words, screening new books, prescribing Setswana literature and translating examination question papers. The Setswana Language Board encountered many problems because authors wrote in their own dialects and the editors rejected the manuscripts since they were not written in their own particular dialect. Radio Bantu was used as the media for promoting the Setswana language. During the same year (1962), a new terminology was compiled and forwarded to the Setswana Language Board to be approved and added to the list of Setswana terminology. In 1964, the committee agreed on the standard orthography, which resulted in the publishing of the Terminology Orthography *no 3*.

2.10.2 Challenges facing the Setswana Language Board

The Setswana Language is comprised seven dialects and the committee was faced with a challenge of having to choose the dialect to be used as a standard orthography.

Comparisons were made of the different dialects. They compared the dialects found in the South and discovered a Sepedi influence, for instance: *mpša* and *mpya* (dog) are derived from the Sepedi dialects. The Eastern dialects were also compared and found to be influenced by the South Sotho dialect, for example; *ntša, ntja* (dog). They finally decided to use *ntšwa* from the Sehurutse dialect and agreed that Sehurutse was to be used as the basis for standard Setswana. When Bophuthatswana became independent, various committees were formed for the publication and evaluation of the manuscripts.

The Setswana Language Board was established under the Department of Education. Other departments such as the Department of Translation and Development (DTD) for Setswana were formulated. According to Ramagoshi (2000), some words in English and Afrikaans were now given new meaning.

Consider the following examples:

Account	>	***Matlotlo***
Accountant General	>	***Mmalamatlotlo kakaretso***
Treasury	>	***letlotlo***
Fund	>	***letlole***

Account	>	***Tshupatlotlo***
Sessional Committee	>	***Komiti ya motlhaakokoano***

(Terminology Orthography no. 4)

2.11 Principles of the Setswana orthography

According to Ramagoshi (2000:60), orthography rules for Setswana have been determined rather by principles (a) to (f), as there were no standards available.

(a) Phonological principle

The phonological principle is suitable for the immigration of loan words. Such words should be adapted to suit the Setswana morphonological rules. Consider the following examples:

Tafel	(*Afrikaans*)	>	*tafola*
Stoel	(*Afrikaans*)	>	*setulo*
Rand	(English)	>	*ranta*
Television	(English)	>	*televisioni*
Fabriek	(*Afrikaans*)	>	*faboriki*
Vye	(*Afrikaans*)	>	*feiye*
Style	(English)	>	*setaele*

(b) Semantic principle

The semantic principle is best suitable in the division of a word and recognized by its pronunciation or written the way it is pronounced. For example, the word *modulasetulo* ‘chairperson’ is written as a single word, but sounds as two separate words. Consider the following examples:

Senotsididi	>	(cold drink or cooldrink)
Modulasetulo	>	(chairperson)
Seyalemowa	>	*Seya* ‘goes’ *le mowa* ‘along the air’ (radio)
Ntlolehalahala	>	(hall)

Compare also the examples written as one word in English, while in Setswana, they must be expressed in several words, as there is no literal translation available

Sekhurumelo sa pitsa	>	(lid)
Ditlamelwana tsa pula	>	(raingarmets)
Kgwedi ya bosupa	>	(seventh month)

(c) Apostrophe and hyphen

Hyphens are used to connect part of speech that expresses certain ideas. The use of the hyphen is prevalent in poetry and should be avoided as much as possible in grammar. Compound words should be written without the hyphen. For example, *Setsayamodumo* (tape recorder) and *lenanathuto* (syllabus) rather than *Setsaya-modumo* and *Lenana-thuto*.

(Terminology and Orthography no. 4)

(d) Principle to facilitate reading

This refers to word divisions, which help to facilitate reading. Words should be written in such a manner as to make reading easier. Written Setswana is divided according to the parts of speech, for example, *yo o buang* (who is talking) is written as three separate words. In this way, the eyes can easily recognise those parts of speech to facilitate reading.

(e) Principle to facilitate writing

The principle facilitates the spelling so that the writer should not make spelling mistakes while writing. It is for that reason why certain words like Sekgowa, Setswana etc. are always written with capital letters, even when they are used as adjectives. This assists the writers not to make spelling mistakes.

(Terminology and Orthography no. 4)

(f) Principle of the history of orthography

The spelling of personal names and surnames written in the old orthography is determined by these principles. For example:

Huhudi	**not**	*Gugudi*
Thabanchu	**not**	*Thabantsho*
Setswana	**not**	*Sechuana*
Kgomo	**not**	*Khomo*

(Terminology and Orthography no. 4)

Principles for the formation of the Setswana orthography rules have previously not been formulated. The only solution to this problem is to include it in the forthcoming Setswana dictionaries.

2.12 Dictionaries

Definition

Leech (1990:204), defines a dictionary as a reference book on the living-room or library shelf. He further describes the dictionary as follows:

> "It is the inbuilt dictionary which every one of us carries around as part of his mental equipment as a speaker of a language".

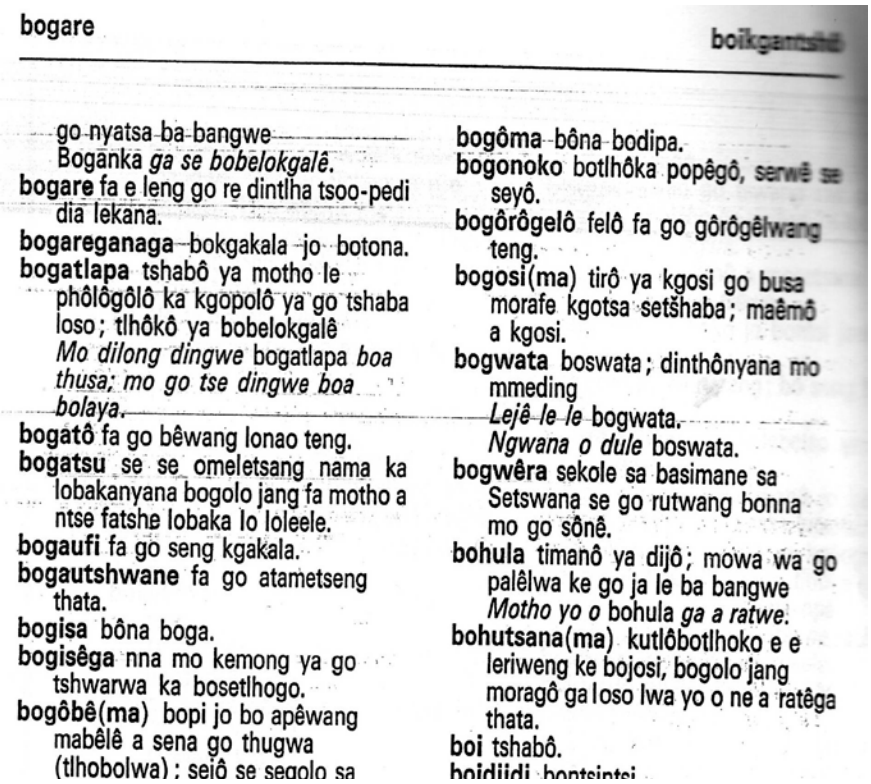
Wikipedia defines a dictionary as a list of words with their definitions, a list of characters with their glyphs, or a list corresponding words in other languages. Many dictionaries include pronunciation information; grammatical information; word derivations, histories, or etymologies; which may use online databases.

There are three types of Setswana dictionaries, namely; (a) monolingual dictionaries, (b) bilingual dictionaries and (c) trilingual dictionaries.

2.12.1 Monolingual Dictionaries

Monolingual dictionaries are designed as an aid to the speakers of the source language i.e. Setswana. Mother tongue speakers as well as the second language speakers who have a broader knowledge of the Setswana language use the dictionary. Two monolingual Setswana dictionaries have been published:

(a) Thanodi ya Setswana ya Dikole (THAND)



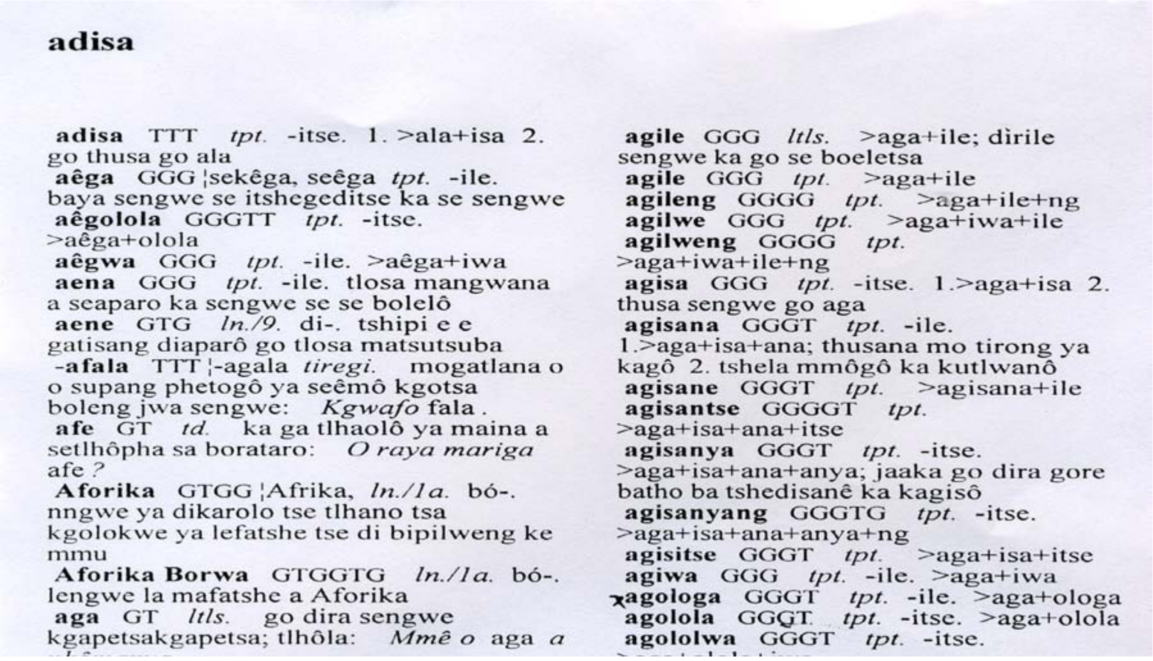
bogare boikgantshê

go nyatsa ba bangwe
Boganka *ga se bobelokgalê.*
bogare fa e leng go re dintlha tsoo-pedi dia lekana.
bogareganaga bokgakala jo botona.
bogatlapa tshabô ya motho le phôlôgôlô ka kgopolô ya go tshaba loso; tlhôkô ya bobelokgalê
Mo dilong dingwe bogatlapa *boa thusa; mo go tse dingwe boa bolaya.*
bogatô fa go bêwang lonao teng.
bogatsu se se omeletsang nama ka lobakanyana bogolo jang fa motho a ntse fatshe lobaka lo loleele.
bogaufi fa go seng kgakala.
bogautshwane fa go atametseng thata.
bogisa bôna boga.
bogisêga nna mo kemong ya go tshwarwa ka bosetlhogo.
bogôbê(ma) bopi jo bo apêwang mabêlê a sena go thugwa (tlhobolwa); seiô se segolo sa

bogôma bôna bodipa.
bogonoko botlhôka popêgô, serwê se seyô.
bogôrôgelô felô fa go gôrôgêlwang teng.
bogosi(ma) tirô ya kgosi go busa morafe kgotsa setšhaba; maêmô a kgosi.
bogwata boswata; dinthônyana mo mmeding
Lejê le le bogwata.
Ngwana o dule boswata.
bogwêra sekole sa basimane sa Setswana se go rutwang bonna mo go sônê.
bohula timanô ya dijô; mowa wa go palêlwa ke go ja le ba bangwe
Motho yo o bohula *ga a ratwe.*
bohutsana(ma) kutlôbotlhoko e e leriweng ke bojosi, bogolo jang moragô ga loso lwa yo o ne a ratêga thata.
boi tshabô.
boidiidi bontsintsi

Kgasa published the first monolingual dictionary THAND in Botswana in 1976. The dictionary comprises of 126 pages. Its main target group are the primary school learners.

(c) Thanodi ya Setswana (THAN)



adisa

adisa TTT *tpt.* -itse. 1. >ala+isa 2. go thusa go ala
aêga GGG ¦sekêga, seêga *tpt.* -ile. baya sengwe se itshegeditse ka se sengwe
aêgolola GGGTT *tpt.* -itse. >aêga+ololа
aêgwa GGG *tpt.* -ile. >aêga+iwa
aena GGG *tpt.* -ile. tlosa mangwana a seaparo ka sengwe se se bolelô
aene GTG *ln./9.* di-. tshipi e e gatisang diaparô go tlosa matsutsuba
-afala TTT ¦-agala *tiregi.* mogatlana o o supang phetogô ya seêmô kgotsa boleng jwa sengwe: *Kgwafo* fala .
afe GT *td.* ka ga tlhaolô ya maina a setlhôpha sa borataro: *O raya mariga* afe *?*
Aforika GTGG ¦Afrika, *ln./1a.* bó-. nngwe ya dikarolo tse tlhano tsa kgolokwe ya lefatshe tse di bipilweng ke mmu
Aforika Borwa GTGGTG *ln./1a.* bó-. lengwe la mafatshe a Aforika
aga GT *ltls.* go dira sengwe kgapetsakgapetsa; tlhôla: *Mmê o* aga *a*

agile GGG *ltls.* >aga+ile; dìrile sengwe ka go se boeletsa
agile GGG *tpt.* >aga+ile
agileng GGGG *tpt.* >aga+ile+ng
agilwe GGG *tpt.* >aga+iwa+ile
agilweng GGGG *tpt.* >aga+iwa+ile+ng
agisa GGG *tpt.* -itse. 1.>aga+isa 2. thusa sengwe go aga
agisana GGGT *tpt.* -ile. 1.>aga+isa+ana; thusana mo tirong ya kagô 2. tshela mmôgô ka kutlwanô
agisane GGGT *tpt.* >agisana+ile
agisantse GGGGT *tpt.* >aga+isa+ana+itse
agisanya GGGT *tpt.* -itse. >aga+isa+ana+anya; jaaka go dira gore batho ba tshedisanê ka kagisô
agisanyang GGGTG *tpt.* -itse. >aga+isa+ana+anya+ng
agisitse GGGT *tpt.* >aga+isa+itse
agiwa GGG *tpt.* -ile. >aga+iwa
agologa GGGT *tpt.* -ile. >aga+ologa
agolola GGGT *tpt.* -itse. >aga+olola
agololwa GGGT *tpt.* -itse.

Kgasa published the second monolingual dictionary THAN in Botswana in collaboration with Tsonope in 1998. THAN is mainly targeted for the secondary school and university students. The dictionary consists of 330 pages.

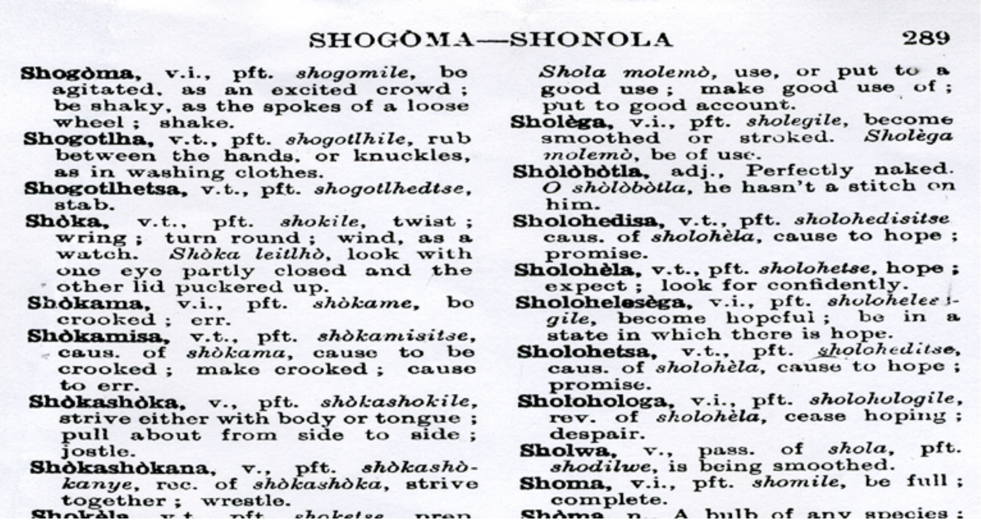
2.12.2 Bilingual Dictionaries

The existing bilingual Setswana dictionaries serve the speakers of both the source and the target language i.e. the Setswana and the English speakers. Two bilingual dictionaries have been published:

a) Secwana-English Dictionary (SED)

b) Setswana-English Dictionary (MSED)

(a) Secwana – English Dictionary (SED)



SHOGÒMA—SHONOLA 289

Shogòma, v.i., pft. *shogomile,* be agitated, as an excited crowd; be shaky, as the spokes of a loose wheel; shake.

Shogotlha, v.t., pft. *shogotlhile,* rub between the hands, or knuckles, as in washing clothes.

Shogotlhetsa, v.t., pft. *shogotlhedtse,* stab.

Shòka, v.t., pft. *shokile,* twist; wring; turn round; wind, as a watch. *Shòka leitlhò,* look with one eye partly closed and the other lid puckered up.

Shòkama, v.i., pft. *shòkame,* be crooked; err.

Shòkamisa, v.t., pft. *shòkamisitse,* caus. of *shòkama,* cause to be crooked; make crooked; cause to err.

Shòkashòka, v., pft. *shòkashokile,* strive either with body or tongue; pull about from side to side; jostle.

Shòkashòkana, v., pft. *shòkashòkanye,* rec. of *shòkashòka,* strive together; wrestle.

Shokèla, v.t., pft. *shoketse,* prep

Shola molemò, use, or put to a good use; make good use of; put to good account.

Sholèga, v.i., pft. *sholegile,* become smoothed or stroked. *Sholèga molemò,* be of use.

Shòlòbòtla, adj., Perfectly naked. *O shòlòbòtla,* he hasn't a stitch on him.

Sholohedisa, v.t., pft. *sholohedisitse* caus. of *sholohèla,* cause to hope; promise.

Sholohèla, v.t., pft. *sholohetse,* hope; expect; look for confidently.

Sholohelesèga, v.i., pft. *sholohelesegile,* become hopeful; be in a state in which there is hope.

Sholohetsa, v.t., pft. *sholoheditse,* caus. of *sholohèla,* cause to hope; promise.

Sholohologa, v.i., pft. *sholohologile,* rev. of *sholohèla,* cease hoping; despair.

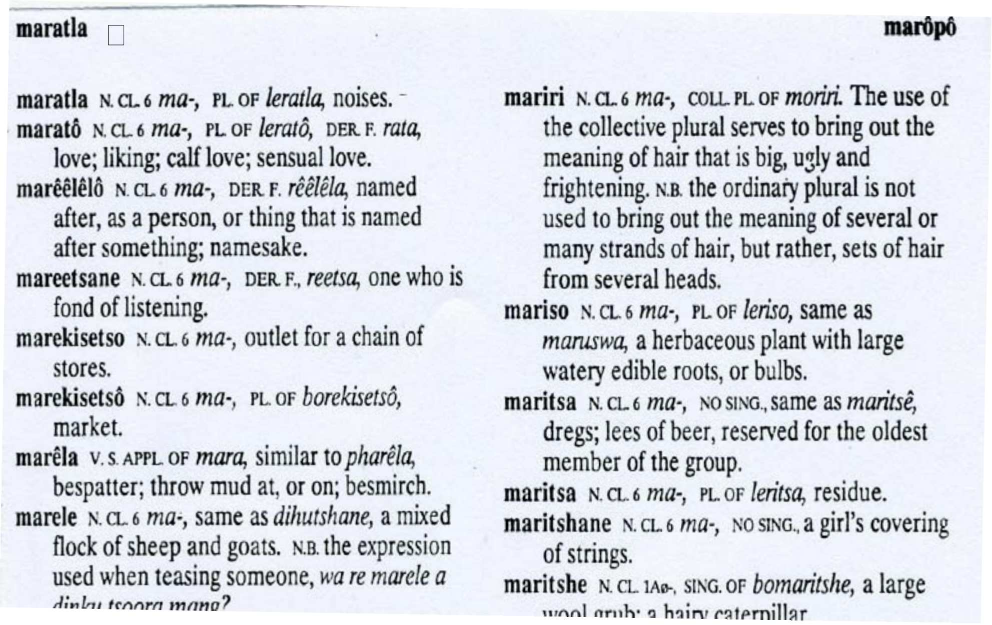
Sholwa, v., pass. of *shola,* pft. *shodilwe,* is being smoothed.

Shoma, v.i., pft. *shomile,* be full; complete.

Shòma, n., A bulb of any species;

The dictionary was published in 1925 by Rev J. Tom Brown. The dictionary was designed to assist the Batswana in the translation of the Bible.

(b) Setswana, English Dictionary (MSED)



maratla □ **marôpô**

maratla N. CL. 6 *ma-*, PL. OF *leratla,* noises.

maratô N. CL. 6 *ma-*, PL. OF *leratô,* DER. F. *rata,* love; liking; calf love; sensual love.

marêêlêlô N. CL. 6 *ma-*, DER. F. *rêêlêla,* named after, as a person, or thing that is named after something; namesake.

mareetsane N. CL. 6 *ma-*, DER. F., *reetsa,* one who is fond of listening.

marekisetso N. CL. 6 *ma-*, outlet for a chain of stores.

marekisetsô N. CL. 6 *ma-*, PL. OF *borekisetsô,* market.

marêla V. S. APPL. OF *mara,* similar to *pharêla,* bespatter; throw mud at, or on; besmirch.

marele N. CL. 6 *ma-*, same as *dihutshane,* a mixed flock of sheep and goats. N.B. the expression used when teasing someone, *wa re marele a dinku tsoora mang?*

mariri N. CL. 6 *ma-*, COLL. PL. OF *moriri.* The use of the collective plural serves to bring out the meaning of hair that is big, ugly and frightening. N.B. the ordinary plural is not used to bring out the meaning of several or many strands of hair, but rather, sets of hair from several heads.

mariso N. CL. 6 *ma-*, PL. OF *leriso,* same as *maruswa,* a herbaceous plant with large watery edible roots, or bulbs.

maritsa N. CL. 6 *ma-*, NO SING., same as *maritsê,* dregs; lees of beer, reserved for the oldest member of the group.

maritsa N. CL. 6 *ma-*, PL. OF *leritsa,* residue.

maritshane N. CL. 6 *ma-*, NO SING., a girl's covering of strings.

maritshe N. CL. 1A*ø-*, SING. OF *bomaritshe,* a large wool grub; a hairy caterpillar

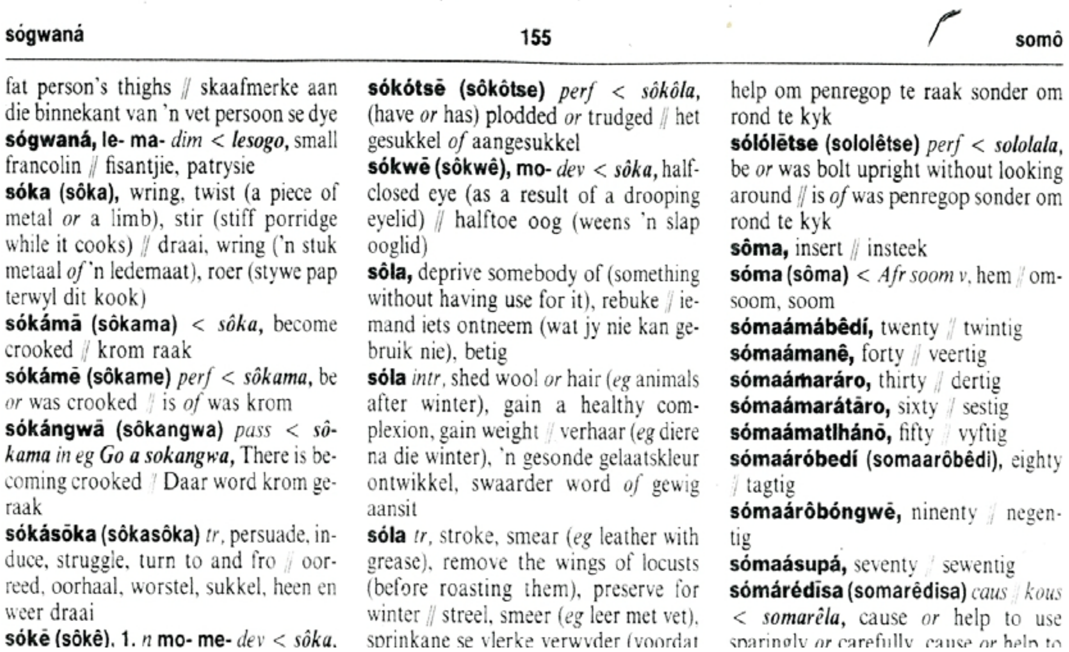
The envisaged revision of *Matumo's* Setswana-*English-Setswana Dictionary* (MSED) is the fourth edition of what is titled since 1993 the *Setswana-English-Setswana dictionary*. The first edition dates back to approximately 1975, the second to 1895, and the third to 1925, entitled SED. The latter was compiled by J. Tom Brown and formed the basis for this dictionary (MSED). It is primarily intended for the English speakers who want to learn the Setswana language visa/verse.

2.12.3 Trilingual Dictionaries

(a) Setswana, English and Afrikaans Dictionary (SEAD)

The *Dikišinare ya Setswana* was published in 1990 by Snyman et al. in South Africa. It is designed to serve all three languages i.e. Setswana, English and Afrikaans. Its target group was the secondary and the university readers.

SEAD



sógwaná 155 **somô**

fat person's thighs // skaafmerke aan die binnekant van 'n vet persoon se dye
sógwaná, le- ma- *dim* < ***lesogo***, small francolin // fisantjie, patrysie
sóka (sôka), wring, twist (a piece of metal *or* a limb), stir (stiff porridge while it cooks) // draai, wring ('n stuk metaal *of* 'n ledemaat), roer (stywe pap terwyl dit kook)
sókámā (sôkama) < ***sôka***, become crooked // krom raak
sókámē (sôkame) *perf* < ***sôkama***, be *or* was crooked // is *of* was krom
sókángwā (sôkangwa) *pass* < ***sôkama*** *in eg* ***Go a sokangwa***, There is becoming crooked // Daar word krom geraak
sókásōka (sôkasôka) *tr*, persuade, induce, struggle, turn to and fro // oorreed, oorhaal, worstel, sukkel, heen en weer draai
sókē (sôkê), 1. *n* **mo- me-** *dev* < ***sôka***,
sókótsē (sôkôtse) *perf* < ***sôkôla***, (have *or* has) plodded *or* trudged // het gesukkel *of* aangesukkel
sókwē (sôkwê), mo- *dev* < ***sôka***, half-closed eye (as a result of a drooping eyelid) // halftoe oog (weens 'n slap ooglid)
sôla, deprive somebody of (something without having use for it), rebuke // iemand iets ontneem (wat jy nie kan gebruik nie), betig
sóla *intr*, shed wool *or* hair (*eg* animals after winter), gain a healthy complexion, gain weight // verhaar (*eg* diere na die winter), 'n gesonde gelaatskleur ontwikkel, swaarder word *of* gewig aansit
sóla *tr*, stroke, smear (*eg* leather with grease), remove the wings of locusts (before roasting them), preserve for winter // streel, smeer (*eg* leer met vet), sprinkane se vlerke verwyder (voordat
help om penregop te raak sonder om rond te kyk
sólólētse (sololêtse) *perf* < ***sololala***, be *or* was bolt upright without looking around // is *of* was penregop sonder om rond te kyk
sôma, insert // insteek
sóma (sôma) < *Afr soom v*, hem // omsoom, soom
sómaámábêdí, twenty // twintig
sómaámanê, forty // veertig
sómaámaráro, thirty // dertig
sómaámarátāro, sixty // sestig
sómaámatlhánō, fifty // vyftig
sómaáróbedí (somaarôbêdi), eighty // tagtig
sómaárôbóngwē, ninenty // negentig
sómaásupá, seventy // sewentig
sómárédīsa (somarêdisa) *caus* // *kous* < ***somarêla***, cause *or* help to use sparingly *or* carefully, cause *or* help to

2.13 Conclusion

In this chapter, we have gained a greater understanding of the origin and the history of the Setswana language. Also how earlier studies have contributed towards documentation and the development of the Setswana language and the challenges the Language Board was faced with in its standardization. The chapter also reveals how the Bantu Education has impacted on the Setswana orthography. Given the available data gathered on the Setswana language, we can conclude that Setswana has been the 'body label' that includes the various Setswana dialects. (See figure 2). An important feature of this chapter is the way maps and diagrams are used to illustrate and show the development of Setswana as a national language and how the South Batswana tribes were distributed. We have seen how the language Board has initiated research and studies aimed at promoting and developing Setswana and the mechanisms identified with the aim of standardising the Setswana language.

The chapter has also shown that language changes offer important evidence about human language — namely that it is rule governed for example, (see 2.8). Important concepts such as analogy and borrowing associated with language change have been defined and discussed. We have also highlighted the changes in the Setswana orthography, spelling rules and abbreviations and how they all impact on the effectiveness of a dictionary. The chapter ends with an overview of the existing Setswana dictionaries which will be dealt in more detail within the forthcoming chapters.

Chapter 3

Compilation of Setswana corpora

3.1. Introduction

This chapter is devoted to the actual compilation of Setswana corpora with special reference to the design of typical corpora such as the Brown corpus and the Longman Lancaster English language corpus. Some of the principles that were used to create the Collins Birmingham University International Language Database (COBUILD) main corpus are also highlighted. COBUILD addresses a number of issues relating to achieving 'balance' and 'representativeness' in the corpus design. These include aspects relating to the 'size' of a corpus. The Setswana organic text corpus and corpus creation is also presented with an explicit, detailed description of how the Setswana oral and written corpus is compiled. A number of techniques and tools used in the corpus analysis for querying the Setswana text corpora are also highlighted.

The corpora are compiled with a few to perform corpus queries mainly in terms of alphabetical and frequency lists, keyness studies and studying keywords in context i.e. the so-called concordance lines. The concept 'keyness' and its two perspectives: 'positive keyness'and 'negative keyness' are defined within the context of the 13 projects from the spoken Setswana corpus. Differences between individuals (narrators), the spoken and the written Setswana corpus are also discussed. Graphs plotting the 'positive keyness' and the 'negative keyness' are included in the discussion. The chapter furthermore highlights the distinction between the Botswana corpus and the South African corpus. Studies to monitor Setswana stability of the growing organic corpora conclude the chapter.

If African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should become an absolute priority according to De Schryver and Prinsloo (2000a, 2000b) and Prinsloo and De Schryver (1999, 2001b). The value of corpora for the compilation of dictionaries is generally accepted in the literature:

> "Contemporary approaches to the investigation of actual language use entail the examination and analysis of collection of different kinds of spoken or written texts, or corpora (the plural of the Latin corpus body). The term corpus linguistics is now used increasingly in the literature, and indeed is found in the titles of a number of influential publications in the field of contemporary linguistic enquiry" (James et al., 1994:4)

3.2 Major (English) electronic corpora in historical perspective

Kennedy (1998) provides a historical overview and evaluates the importance of a corpus-based approach to a dictionary. He gives a detailed presentation of the major English corpora, provides taxonomies to organize the field and distinguishes the first generation Brown corpus and Lancaster-Oslo/Bergen (LOB) corpus from the second generation.

3.2.1 COBUILD main corpus

(a) Aims and objectives

According to Kennedy (1998), the first COBUILD project started in 1980 with a particular commercial research purpose and development project to produce corpus-based dictionaries, grammars and language teaching courses.

This means that the dictionary is based on a 'corpus', which is a collection of British and American newspapers, books, TV programmes, real-life conversations including textbooks, novels, guides, magazines and Websites. The corpus has been automatically word-class tagged and a corpus of 200 million words has been parsed.

The COBUILD is updated and added to on a regular basis to ensure that this resource is as up to date and comprehensive as possible. According to Kennedy (1998:3-4), the COBUILD corpus is designed for general linguistic purposes, that is, to answer questions at various linguistic levels on the prosody or lexis grammar and discourse patterns or pragmatics of the language. Allen (2006:2) regards the COBUILD corpus, which consists of 56 million words of written and spoken text, as the bank of English.

(b) Sampling principle

Some of the principles that were applied in the creation of COBUILD were enumerated by Renouf (1987:2-5). The text consists of 7.3 million words, 25% spoken text, general rather than technical language from 1960 onwards, naturally occurring text as well as writing and speech produced by adults aged 16 and over. The spoken text included transcripts of radio broadcasts, university archives of oral interviews and lectures. Written texts were chosen from widely read works (excluding poetry) and the authorship was 25% female. Newspaper and journalistic texts were included (see Prinsloo and De Schryver, 2000:8)

3.2.2 LOB corpus (Lancaster Oslo Bergen)

The Lancaster Oslo-Bergen (LOB) corpus is a synchronic corpus of approximately one million words representative of written English text. The Brown corpus consists of 500 samples of 2, 000 words each, taken from different books, newspapers, etc. (See Table 1 below).

(a) Sampling principle

The overall method used in sampling is to randomly select titles from bibliographical sources. For each text extract selected, a check was made as to whether the author was British, although this could not always be established. Texts published by non-British authors were excluded.

In selecting text extracts, an attempt was made to limit the amount of dialogue to 50% or less although this was not always feasible. The modification of purely random sampling was used extensively in compiling the categories of newspaper prose and the selection of newspapers was weighted in favour of the national press. Similarly, major periodicals were favoured over less important ones.

Table 1: Basic composition of Brown and LOB corpora

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

(Johansson and Hofland, 1989:2)

3.2.3 Longman Lancaster English language corpus

Texts were collected in two ways i.e. a selective half was chosen through a mixture of pragmatic measure to gather a broad range of objectively defined ‘document types’; and a microcosmic half was obtained by randomly selecting books. (See Figure 6 on the next page.) The use of ‘document types’ was introduced by Michael Rundell and was defined as ‘text from a particular subject area, together with a cluster [of] relatively identifiable features such as time region; medium and level’. (Summers, 1993:192). Broad subject areas were then adopted, namely natural and pure science

(6,0%), applied science (4,3%), social science (14,1%); world affairs (10,4%); commerce and finance (4,4%); arts (7,9%), belief and thought (4,7%); leisure (5,7%), fiction (40,0%) as well as poetry, drama or humour (2,3%).

Summers (1993:192,193) noticed that the absence of the spoken corpus was essentially topic driven rather than genre driven, as cited by Prinsloo and De Schryver (2000a) and one can notice the absence of spoken sources. According to Summers (1993:184), the importance of the spoken language was overestimated and there was insufficient written material. However, a large body of speech was adequately captured electronically.

Irrespective of the arguments about the inclusion of spoken texts, one has to keep in mind the exact purpose for which the corpus is intended. Moon (1998:353) suggests that the constraints of the conventional dictionary “make it difficult if not impossible to deal with the distinguishing features of spoken language properly and fully”.

In fields such as speech-processing technology, corpora exclusively consist of oral material in spite of the huge practical, technical, financial and ethical problems associated with the acquisition of spoken data. Moon, (1998:348) also confirms the financial aspect as the real stumbling block for the erection of large oral sub-corpora.

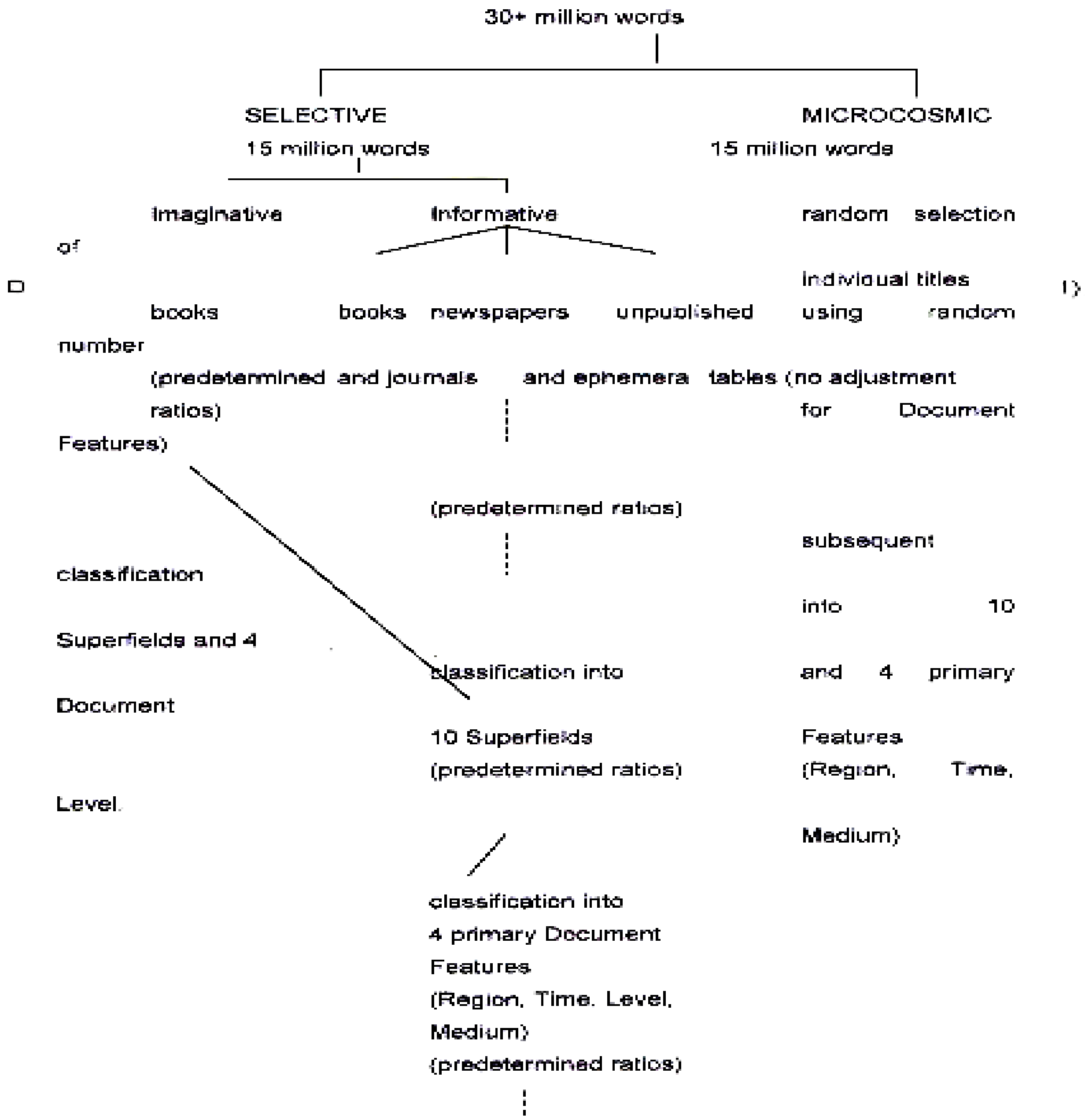

30+ million words
SELECTIVE
15 million words
MICROCOSMIC
15 million words
imaginative
informative
random selection of individual titles using random number tables (no adjustment for Document Features)
books (predetermined ratios)
books and journals
newspapers
unpublished and ephemera
(predetermined ratios)
subsequent classification into 10 Superfields and 4 primary Document Features (Region, Time, Level, Medium)
classification into 10 Superfields (predetermined ratios)
classification into 4 primary Document Features (Region, Time, Level, Medium) (predetermined ratios)

Figure 6: Longman/Lancaster English language corpus – current structure

Atkins (1997; oral communication at Salex 97) has an interesting approach to the concept ‘organic corpora’. According to Atkins, a corpus builder first attempts to create a representative corpus. Then this corpus should be used and analyzed and its strengths and weaknesses identified and reported. It is enhanced by the addition or deletion of material and the circle repeated continually. Furthermore, one should not try to make a comprehensive and watertight listing … rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language … In their ten years' experience as a team of analyzing corpus material for lexicographic purposes, they have found any corpus however unbalanced to be a source of information and indeed inspiration. Important for the lexicographers, is knowing that their corpus is unbalanced. Organic corpora are of specific significance to African lexicographers as written sources are scarce.

3.3 Issues ‘size’, ‘balance’, ‘representative’ and ‘organic corpora’

According to Biber (1993), there is a vast literature on the issue of representativeness of corpora.

Shannon (1948:5) states:

> "A compact way of representing a collection of documents is by means of a frequency list, where each word is associated with the number of times it occurred in the collection. The representation defines a simple language model, a stochastic approximation to the language used in the collection, i.e. an 'oath order' word model or a unigran model. As the model's complexity increases its approximation to the target language improves"

Ciaramita and Baroni (2006:21), define balance in terms of the set of biased corpora that one compares the target corpus against. They state:

> "Assuming that our measure of unbiasedness/balance is appropriate, all it tells us is that a certain corpus is more or less biased than another corpora we compared them against e.g. the corpus built with the mid frequency seeds is less biased than the others with respect to corpora that represent 10 broad topic-based WordNet categories."

Kennedy (1998:20, 52, 62) states that the sample is at best a rough approximation to representativeness, given the vast universe of discourse … the issue is really representative of what? In light of the perspectives on variation offered by several decades of research in discourse analysis and socio-linguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres.

Other linguists such as Wilson (1996) also present overviews of the theory and practice of corpus linguists and further emphasize the key factors in a corpus based approach such as sampling, representative, size, balance etc. based on the Brown and LOB corpora.

According to Kilgarriff and Grefenstette (2003:1), corpora are not meant to represent a specific sub-language but the language as a whole. Baroni and Bernadini (2004:14) state:

> "We must try to construct a balanced corpus by selecting appropriately balanced query terms, e.g. using random terms extracted from an available balanced corpus. In order to build specialized domain corpora, we will have to use biased query terms from the appropriate domain."

Linguists disagree whether a corpus should try to be balanced or representative. It seems as if a corpus will never be balanced because of the many parameters and never be truly representative of all language usage either, as it is impossible to define the population. All the compiler can do is to strive to come as close to the ideal as possible.

3.3.1 Representativeness and balanced samples

Notwithstanding the fundamental shortcomings even in defining the concepts balanced and representative, corpus compilers generally strive to achieve these goals. Both the Brown corpus and the British National Corpus (BNC) have made serious attempts to be balanced and representative. Brown consists of 500 samples with 2,000 words each, taken from different books, newspapers, etc. The BNC contains more than 4,000 documents of widely different sizes and a collection of 4,000 short books from the library. A balanced corpus is representative of the relevant sub-language, because it contains material from all the different genres. For example, the BNC contains slightly more than 10% of spoken materials. If BNC frequencies are taken to be representative of modern British English, there is an implicit assumption that only 10% of the output of British speakers consists of speech while the remaining 90% are produced in writing.

According to Prinsloo (2004: 33), what is important for lexicographic work in South Africa, is that corpus compilers should be sensitive to all these aspects i.e. to build as far as possible, corpora that are big enough, well balanced and representative so that valid conclusions for lexicographic purposes can be drawn.

3.4 Compilation of the Setswana corpus

The aim was to build a corpus for the Setswana language from both spoken and written texts. Written texts were chosen from widely read works and poetry, educational newsletters, former Bophuthatswana advertisements and journalistic texts. The recordings were transcribed manually on a word processor using the available orthography and instantly saved as text files. Still one will do well to keep Kennedy's observation in mind:

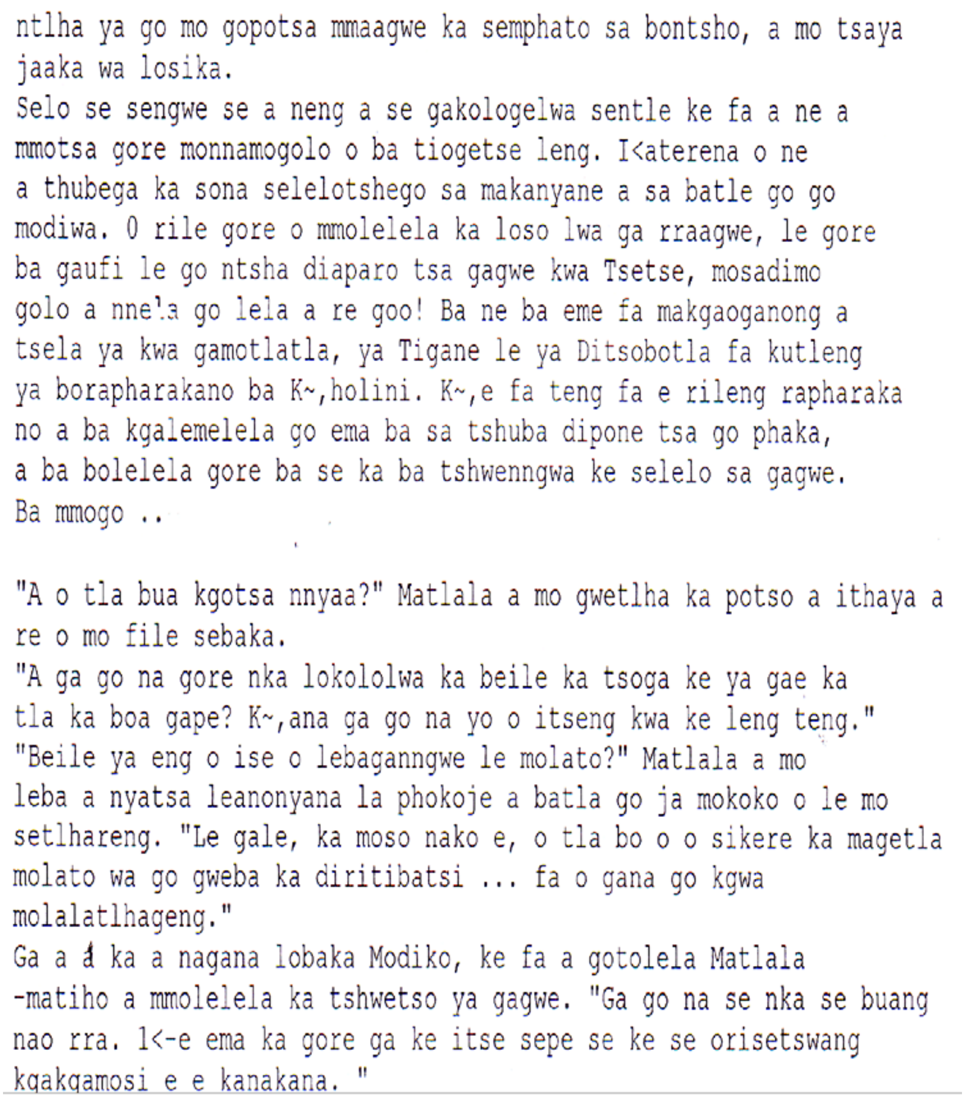
> "A transcription is an imperfect written approximation of a speech event, which exists initially as a dance of air molecules. The level of delicacy or

> amount of details in a transcription is … related to the use which the transcription will be put" (Kennedy, 1998:82).

According to Prinsloo and De Schryver, (2000:11), there are three ways of entering written materials into computer files: (a) electronic transfer, (b) (re)keyboarding and (c) scanning. The Setswana text collection of written materials was scanned using (OCR) software. A number of problems were encountered when running (OCR) software and most of them had to be rectified manually. The following examples of misreading occurred consistently across most sources during the creation of the Setswana corpora. The following quotation taken from the Setswana oral projects reflects the typical (OCR) output and Table 2 typical misreadings.

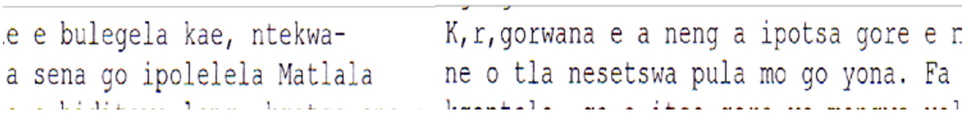
Typical OCR output

“



ntlha ya go mo gopotsa mmaagwe ka semphato sa bontsho, a mo tsaya
jaaka wa losika.
Selo se sengwe se a neng a se gakologelwa sentle ke fa a ne a
mmotsa gore monnamogolo o ba tiogetse leng. I<aterena o ne
a thubega ka sona selelotshego sa makanyane a sa batle go go
modiwa. 0 rile gore o mmolelela ka loso lwa ga rraagwe, le gore
ba gaufi le go ntsha diaparo tsa gagwe kwa Tsetse, mosadimo
golo a nne'a go lela a re goo! Ba ne ba eme fa makgaoganong a
tsela ya kwa gamotlatla, ya Tigane le ya Ditsobotla fa kutleng
ya borapharakano ba K~,holini. K~,e fa teng fa e rileng rapharaka
no a ba kgalemelela go ema ba sa tshuba dipone tsa go phaka,
a ba bolelela gore ba se ka ba tshwenngwa ke selelo sa gagwe.
Ba mmogo ..

"A o tla bua kgotsa nnyaa?" Matlala a mo gwetlha ka potso a ithaya a
re o mo file sebaka.
"A ga go na gore nka lokololwa ka beile ka tsoga ke ya gae ka
tla ka boa gape? K~,ana ga go na yo o itseng kwa ke leng teng."
"Beile ya eng o ise o lebaganngwe le molato?" Matlala a mo
leba a nyatsa leanonyana la phokoje a batla go ja mokoko o le mo
setlhareng. "Le gale, ka moso nako e, o tla bo o o sikere ka magetla
molato wa go gweba ka diritibatsi ... fa o gana go kgwa
molalatlhageng."
Ga a á ka a nagana lobaka Modiko, ke fa a gotolela Matlala
-matiho a mmolelela ka tshwetso ya gagwe. "Ga go na se nka se buang
nao rra. 1<-e ema ka gore ga ke itse sepe se ke se orisetswang
kgakgamosi e e kanakana. "



.e e bulegela kae, ntekwa-
a sena go ipolelela Matlala

K,r,gorwana e a neng a ipotsa gore e r
ne o tla nesetswa pula mo go yona. Fa

The symbol ‘← →’ indicates that the scanning error occurred in both directions e.g. ‘c ← → o’ means an o is incorrectly scanned as ‘c’ and ‘c’ is correctly scanned as ‘o’. Consider the examples in Table 2 below:

Table 2: Typical OCR misreading of signs

*instead of	*I < → K	*n I, - o →nk	*c → e
*lr → k	*k111 → k	*I,, → k	*nc → ne
*I → ko	*I ~c → k	*o → s	*i → di
*I11 → I	*IIc → k	*I-, → k	*I- → e
*Iıı → k	*Ic → k	*o → o	*j~a → Ia
*I → k	*I ~c → k	*c← →o	*I~c → k

In Setswana, some letters of the alphabet are not used. For example, ‘c’ and ‘q’. Since it is discovered that a large number of OCR errors are consistent and not necessarily OCR trainable, one is advised to keep track of recurring miscannings and use the straightforward ‘search and replace’ function of a word processor to perform corrections. In addition spellcheckers can also be used to clean recurring miscannings. Subsequent checking and recognized files may be required if high levels of accuracy are called for.

3.4.1 Text encoding

Text encoding activities are forthcoming and briefly outlined here. According to Prinsloo and de Schryver (2000:13) text encoding can consist of any combination of (a) word tokenisation, (b) part-of-speech tagging, (c) lemmatisation and (d) syntactic parsing.

(a) Word tokenisation

Many African languages, particularly Setswana, contain very few word delimiters as they have a disjunctive orthography. Segmenting a text containing conjunctively written words into freestanding words is known as "word tokenisation" (Mills, 1998:213, 215).

(b) Part-of-speech tagging

This concept refers to the assigning of a word class to all the words in a text by means of the grammatical data for lemmatisation, parsing or advanced concordance.

(c) Lemmatisation

According to Hartmann and James (1998:83), lemmatisation should be understood as 'the reduction of a paradigm of variant word forms to canonical form' e.g. the inflected forms (***-ng***, ***-el***, ***-ga***) of Setswana locatives verbs in the following examples:

Example 1

Locative *(-ng)*

Lemma		
sekolo(school)	*seko**long***	**(to school)**
naga (veld)	*nag**eng***	**(to the veld)**

Verbs

Lemma

reka **(buy)**	*reke**la***, *rekele**la***, *reke**ga***, *reki**sa***, *rekise**tsa***.
bofa **(tie)**	*bofe**ga***, *bofele**la** a*, *bofolo**la***, *bofolo**ga***, *bofi**sa***

(c) Syntactic parsing

According to Kennedy, (1998:21) corpora can also be parsed to show the sentence structure and the function in the sentences of the different word classes. Consider the following example:

Example 2

Monna	*yo o*	***jang***
The man	**who**	**is eating**
Noun	relative concord	verb stem with relative *-ng*

For the purpose of this study the aim was simply to compile raw corpora.

3.4.2 Querying text corpora

According to Prinsloo and De Schryver (2000:15), corpora are of no use without powerful corpus query tools as a minimum requirement. Such tools must be:

- able to deal with huge numbers of text files
- handle files stored in plain texts as well as in mark-up format
- calculate basic statistics
- present alphabetical and frequency word lists
- provide concordance lines for lexicographic purposes.

Corpora are mainly queried to obtain/generate alphabetical word lists; frequency lists reflecting overall and comparative counts or contexts reflecting words in context. There are quite a number of software packages available to perform these tasks such as corpus Bench from Denmark; monoConc (www. Michaelbarlow.

Com/viz.html) from the United States of America (www.lexically. net/ download/ version4/htm/index.htm), WordSmith Tools from England. (www.lexically.net/ wordsmith/corpus _ linguistics _ links/Wilkinson.doc). (See also Rundell, 1996:16–19; Kennedy, 1998:259–267). For this study WordSmith Tools was selected.

The Word list function of WordSmith Tools generates word frequency and alphabetical lists and indicates the number of types and tokens. According to Lancashire (1993:293), the frequency count list will assist the lexicographer to identify the most frequently used words and to explain how different genres and sub-genres influence their use.

3.4.3 Alphabetical word list and frequency word list

Table 3: Frequency word list reflecting overall counts for the 100 most frequently used words in the South African Setswana text corpus

Rank	**Word**	**Frequency**
1.	*A*	305,598
2.	*GO*	162,791
3.	*LE*	157,695
4.	*E*	142,852
5.	*O*	130,611
6.	*BA*	129,940
7.	*KA*	126,974
8.	*KE*	112,196
9.	*YA*	89,307
10.	*MO*	88,179
11.	*GA*	70,194
12.	*FA*	66,508
13.	*RE*	65,782
14.	*SE*	63,062
51.	*KGOSI*	6,889
52.	*JALO*	6,886
53.	*MORAGO*	6,777
54.	*GONNE*	6,601
55.	*I*	6,546
56.	*THATA*	6,522
57.	*MONNA*	6,484
58.	*ENG*	6,457
59.	*UTLWA*	5,926
60.	*SETSE*	5,707
61.	*MORENA*	5,688
62.	*WENA*	5,662
63.	*RONA*	5,569
64.	*RILE*	5,493

15.	*NE*	49,119
16.	*DI*	49,009
17.	*WA*	44,516
18.	*GORE*	44,080
19.	*SA*	37,761
20.	*TSA*	32,849
21.	*KWA*	31,842
22.	*TLA*	29,874
23.	*MME*	27,303
24.	*TSE*	25,076
25.	*BO*	22,938
26.	*LA*	20,863
27.	*GAGWE*	20,385
28.	*YO*	19,308
29.	*NNA*	19,259
30.	*BONA*	18,773
31.	*LO*	15,298
32.	*FELA*	14,566
33.	*NA*	14,108
34.	*NTSE*	12,148
35.	*ITSE*	11,559
36.	*JAAKA*	11,259
37.	*JWA*	11,210
38.	*NENG*	10,831
39.	*MOTHO*	10,691
40.	*ME*	10,028
41.	*BONE*	9,891
42.	*BATHO*	9,749
43.	*MONGWE*	8,449
44.	*GAGO*	8,130
45.	*TENG*	8,012
46.	*BUA*	7,519
47.	*PELE*	7,282
48.	*TSWA*	7,220
49.	*ENE*	7,066
50.	*DIRA*	7,053

65.	*NTLHA*	5,312
66.	*YONA*	5,306
67.	*PELO*	5,052
68.	*NGWANA*	5,025
69.	*BATLA*	5,024
70.	*TSENA*	4,941
71.	*NAKO*	4,919
72.	*GAPE*	4,864
73.	*KGOTSA*	4,854
74.	*LETSATSI*	4,838
75.	*BANA*	4,788
76.	*MOSADI*	4,505
77.	*JAANONG*	4,497
78.	*TSAYA*	4,464
79.	*TSAMAYA*	4,436
80.	*TLE*	4,365
81.	*MAFOKO*	4,294
82.	*ENA*	4,278
83.	*SENGWE*	4,247
84.	*BILE*	4,132
85.	*SENTLE*	4,119
86.	*DILO*	4,057
87.	*JANG*	4,011
88.	*GODIMO*	4,010
89.	*KANA*	3,983
90.	*RATA*	3,932
91.	*MATLHO*	3,906
92.	*RAYA*	3,896
93.	*BE*	3,881
94.	*KAE*	3,821
95.	*MODIMO*	3,817
96.	*JAANA*	3,767
97.	*NNGWE*	3,707
98.	*JO*	3,690
99.	*LWA*	3,687
100.	*TSOTLHE*	3,679

Table 3 is a word list containing the most frequently used words and their overall counts i.e. number of occurrences in the corpus in column 3. Column 1 reflects the ranks of each word e.g. the most frequently used orthographic word is A with rank 1. A huge proportion of words occurs once only and is called hapax legomena. Most of the top 100 words are closed-set items, a weft of prepositions, determiners, pronouns, conjunctions; whose role is mostly to glue texts together by supplying grammatical information to a lexical wrap of nouns, verbs, adjectives and adverbs.

Scott and Tribble (1996) define a word list essentially as a list of word types. A word list program goes through a text or a set of texts and reduces all repeated tokens to types (counted once) together with its frequency - hence, the important distinction between ‘types’ (the different words in a corpus) and ‘tokens’ (the running words in a corpus).

3.5 Compilation of a Setswana oral corpus

It is unfortunate that most corpora around the world lack sufficient data from spoken sources. The reason for this is that there are many logistical problems and ethical factors involved in the collection of spoken data and the collection process is much more time consuming and expensive. The oral data was drawn from 60-minute tape recordings of individual interviews. The themes included Setswana poems and praises, wedding and birthday celebrations, including radio broadcasts. Oral data can pinpoint words, which tend to be used more frequently in oral versus written communication. Brief references will be made to keyness which will be dealt with in detail in this chapter.

3.5.1 Keyness definition

Keyness is defined as frequency, thus the word is regarded as key if its frequency is much more or much less than it is expected in comparison to its reference corpus. Williams (1983:14–15), defines keyness as:

> "… strong, difficult and persuasive words in everyday usage …
> common in descriptions of wider areas of thought and experience …
> they are significant, binding words in certain activities and their
> interpretation; they are significant, indicative words in certain forms of
> thought".

Andor (1989) considers keyness words as ‘items that could function as keywords in a passage or chain of words where they are dominant’.

In this study a word can be regarded as ‘key’ if, in a specific corpus, it is used much more (positive key) or much less (negative key) than it is expected in terms of a more general (usually bigger) corpus of the language.

The flowchart in Figure 7 below will now present the recording and processing of the Setswana oral corpus.

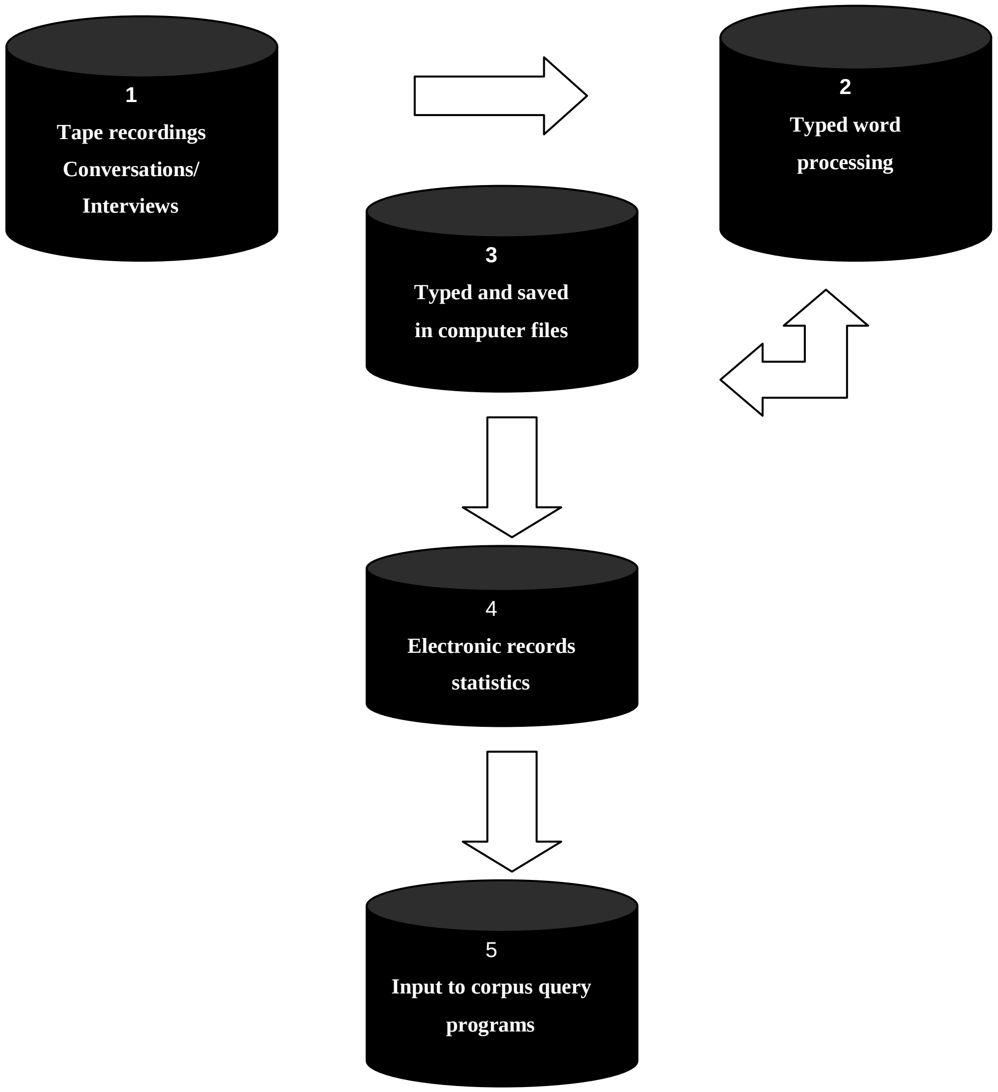


Figure 7: Flow chart representing oral corpus recordings and processing

The tape recordings consist of 13 samples chosen from randomly selected persons i.e. male and female professional and non-professional adults, ranging from 20years to 60years of age. (See Appendix 2)

The conversations, interviews, poems, praise and wedding celebrations were recorded and then typed onto a computer. The semi-structured interviews of each candidate include the following basic questions:

Aneela ka bokhutswane ka tsa botshelo jwa gago. (Briefly tell about your life history)
Bapisa botshelo jwa gago le botshelo jwa segompieno. (Compare your life style with the present life style.)
Ke eng se o eletsang go bona se fedisiwa kgotsa o sa se rate? (In your own opinion, which things do you consider most irrelevant or you dislike in the modern way of life?)
A puo le setso ya Setswana *di a somarelwa kgotsa di anyelela*? (Do you consider the language and culture of Setswana conserved or destroyed?)

The exact oral collections will now be briefly analysed with identification of comments on certain lexical items used in the oral communication.

3.6 Summary of the Setswana oral projects

A brief summary of the narrator's attitudes will be given as well as highlighting peculiarities of oral communication in comparison to the written corpus.

Project 1: An elderly lady.
The narrator was born in 1914 in the village of Hebron near Pretoria. She is a skillful person with lots of experience. She perceives youths of the modern society as troublemakers. The narrator is very proud about her habitat. She uses Setswana with pride and expresses the way of life of that particular environment. She is a person who is rooted to her present surroundings and has no intention to relocate. The narrator is old, lacks education or knowledge and confuses dates and places. Her language identifies her lifestyle and her heritage. She depends on natural resources for survival and most of her terminology is based on her environment.

The narrator is sceptical about modern life and often criticizes it. She was raised under the influence of Afrikaans and English and her beliefs are mixed with other religions that were prevalent in her youth.

The keywords *ntate* (father) (rank 25) in the Setswana oral corpus is repeated 19 times and the days of the week, such as *Sontaga* (Sunday) and *Satertaga* (Saturday) as opposed to *rre* (father) (rank 131) are used with a frequency of 3,659. *Latshipi* (Sunday) and *Lamatlhatso* (Saturday) respectfully, are used far less often in spoken and formal conversations than in written texts and meetings. Most Setswana dialects tend to prefer *ntate* to *rre* even in formal gatherings. These words are not standard Setswana words, thus they fall below the filter threshold in the word list. They thus indicate negative keyness as they are used less often than expected in the general corpus.

Project 2: A male educator at a particular School.

The narrator was born in 13th 1964 in the village of Warmbad. He is the first born and the only son in the family of five children. He proved to be an intelligent person as he skipped two standards (grades) in one year. He grew up under difficult conditions. Although he was ambitious, he was unable to achieve his goals owing to unfavourable circumstances. The narrator encountered challenges that were beyond his comprehension at an early age. He is pleased by every little achievement that he makes. He retains his youthful ambitions and is prepared to take on more challenges. His language is mainly connected to his tradition or culture. According to the life he has led, he is supposed to at least align himself with certain cultural or traditional events. He displays no interest in identifying with the past but concentrates on the present. His words are closely attached to the conditions and the environment in which he lives.

He uses borrowed words because he lacks alternative vocabulary. The narrator was deprived of opportunities because of circumstances beyond his control. In spite of all the hardships he has experienced, he could venture into the unknown and still make a positive contribution.

The keywords like *koo* (there) (rank 380) with a frequency of 1,178, *foo* (there) (rank 200) with a frequency of 2,328 and *yoo* (that one) (rank 370) with a frequency of

1,223 are repeated more often than expected in the text. The repetition of words such as *foo*, *koo* and *yoo* signals an influence from the Serolong dialect.

Project 3: A certain farmer in a village.

The narrator was born in Tawana near Hammanskraal. He managed to pass Standard 4 as his highest level of education. His parents then moved to Warmbad where they worked on the farm. Unfortunately they were badly treated by the farm owner and they decided to move to Ga-Rankuwa in search for better employment. The narrator was raised with good morals and parental support as is perceived by the good words he uses when portraying his well-being. As farmers, his parents adopted farm life and had a fear of challenges. The old system is partly to blame, on the other hand, there were people who moved away from farm life and prospered elsewhere. Farm life deprived him of the opportunity to be educated but he went to school until Standard four. His hatred of the past still exists, but he realised that he cannot live in the past and should rather concentrate on the future. His language is not authentic, but is built on other dialects and borrowed words. He has lost his cultural identity and often uses loanwords like *mara* (*maar*/but) instead of *lemororo* (although) and other *Sotho*-related words.

He has a very strong character, despite the hard life he has led; he does not give up his ambitions. He shows appreciation and forgiveness and therefore has a healthy personality. He passes on the good morals and respect he was taught by his parents to his children. He is self-determined and focussed.

The keywords *ra* (we) (rank 80) with a frequency count of 1002 and *re* (rank 14) with a frequency count of 12447 are repeated more often in the text than expected. Thus these words signal togetherness as the narrator always mentions the word (we). Although the narrator is too general in his speech, his words are among the most highly frequently used in the general corpus and are regarded as positive keys since they best describe the text.

Project 4: Programme director at the graduation ceremony.

The programme director is the person responsible at a radio station for day-to-day management of Setswana programmes. She has a lot of service experience in this field. She is flexible, thus the guests thoroughly enjoyed themselves as they got a chance to participate in a variety of activities such as music and praise. The programme director uses the language that is highly influenced by township and modern cultures. As people from different walks of life are invited to a graduation ceremony, he is catering for most listeners. He is a Setswana speaker by nature, but has adapted to the township language. As a programme director, he focuses on what people would like to hear and understand. The narrator moved away from his native Setswana mainly because of language influence from cultural infiltration and his environment. He does not identify himself with the Setswana rituals because the graduation ceremony is of a modern practice.

The programme director cannot be blamed for using a lot of borrowed words such as party instead of *moletlo*, pressure instead of *kgatelelo*, *khamera* instead of *setsaya-ditshwantsho etc*., because he wants to reach his audience, who consist of youth as well as adults. He also uses motivational language while encouraging introspection from the graduates. It is evident that the ceremony is modern because electronic equipments were frequently used.

The use of exclamations such as *eeee, g-g, so, a-aa* and *erg* are repeated to signal more relaxed and informal gatherings. These keywords consists of what Scott (1999) calls the 'aboutness' variety (words that tell us about the genre of the corpus). These words tend to be overlooked as keywords simply because they do not occur often enough to make a sufficient impact, but taken as a cumulative whole, would actually appear as key.

Project 5: A lecturer at a particular university.

The narrator is a black South African who speaks other African languages and also respects other African cultures. He grew up in Botswana, taught Setswana first language at various institutions and has a very good command of the Setswana language. His lecture is based on the differences and similarities that exist between language and culture. He is well travelled, understands various cultures and outlines the uniqueness of all traditions by their attire and use of language. The narrator singled out all family members according to their portfolio's and responsibilities. He understands all black rituals, their function and preparation according to each special occasion. He showed how different cultures cater for unrelated children and less fortunate adults to make them feel acceptable, for example, donations are given to adults to help them fulfil their obligations as parents to their children.

The narrator understands the Setswana language structure and attaches meaningful interpretation to his words to benefit readers. He appreciates how this language enriches the lives of both the Batswana adults and youths. He identified certain hidden meanings that could only be understood by real Batswana people and he shows a lot of respect for his language.

The narrator reveals the authentic life styles of the Batswana people, who based their prosperity on their nature and animals. This is reflected in their language idioms and expressions, which reflect their reference for and dependence on their animals and nature. The narrator preserves his environment in order to benefit from it and shows how the environment contributes to certain life styles.

The keywords *kgomo* (cow) (rank 210) is repeated 36 times in a text to signal the importance of a cow in the life of the Batswana people and *jaaka* (for example) (rank 36) with a frequency of 11,559 is repeated 13 times. The narrator attaches meaning to the past as he speaks about cows and often plays around with repetition of words. The repetition of words such as *lenweenwee* and *bonweenwee* (mistrust) etc., signal the importance or value of his speech. Many of these words are characteristics of the language of time of farming and also resemble certain features. These words are used less often in the spoken than in the written texts. As a result they do not appear within the 100 most frequently used words in the word list of the oral corpus.

Project 6: A resident in Makapanstad area.

The narrator was born in 1938 in Lady Selborne (Pretoria). At seven years when attending grade 4, his parents decided to move to East Lynne. He was brought up by his aunt under very strict conditions. He therefore learned to be responsible at an early age. Due to lack of finance, he was forced to seek employment at the Pretoria Market. He did not enjoy the warmth of a family owing to his socio-economic background. Apart from all the difficulties he was faced with, he persisted in achieving his dreams. Irrespective of the environment in which he grew up, he was not negatively influenced and chose not to succumb to circumstances that were destructive and threatening to his well-being. The narrator is capable of making things happen irrespective of the challenges he was faced with. He is either voluntarily or forcefully aligned with the correct attitudes and values of the communities in which he finds himself. The narrator has a strong character and faces challenges with, willpower faith and hope.

The keywords like *ko* (rank 364) frequency of 1,264, *moo* (there) (rank 186) frequency of 2,457, *yoo* (that one) (rank 370) frequency of 1,223 and *koo* (there) (rank 308) frequency of 1,178 occur more frequently in the oral text than it is expected. Most of this words are demonstrative nouns and are in one way another influenced by the Sekgatla dialect that is spoken around the area of Makapanstad.

Project 7: An educator in Makapanstad.

The narrator was born in Bosplaas in the Makapanstad area. He is well educated and has taught Setswana first language in secondary schools, colleges as well as Universities. He is the author of many Setswana novels and a grade 12 Setswana examiner for more than thirty years. He also participated in structures such as PanSALB and the Setswana Language Board.

The narrator notices various Setswana language structures. The Setswana language has many dialects that have been influenced by other various cultures. Although the narrator uses the standardized form of language, she sometimes uses borrowed or loanwords. She is selfish about the Setswana language. The narrator is disturbed by the fact that the Batswana people today have to compromise their language because of technology and changes. The problem is that the Setswana language lacks scientific and modern technology terminology. According to the speaker, the Setswana spoken by the youth is not recommended since it is misleading and incorrect. She criticises technology and encourages the Batswana people not to be depended on loanwords or borrowed words but rather to formulate their own vocabulary. The narrator further recommends that the Batswana people should go back to their roots. For example, use of traditional medicines and their names should be protected and reserved.

“*Kgomo e ne e tshasiwa ka mafura a tlou, tau di ne di e tshaba*”. This information is busy disappearing since most people are migrating to cities. The keyword *re* (we) rank 14 is repeated 21 times and occurs much more frequently than it is expected in the spoken corpus. The word *re* (we) is a pronoun used by the narrator to express emotions of togetherness and encourages other people to start new terminologies to protect the Setswana culture.

Project 8: Lesson presentation by a grade 12 educator.

The narrator is a highly educated person with expertise in Setswana literature and a lecturer at a University. He is the author of many Setswana literature books. The Setswana months were identified by the natural phenomena that surround their daily lives. Take for example, *Seetebosigo* (June month). During this month of the year it is very cold and people are warned not to visit each other since blankets are very scarce. The meanings were unique because the Batswana people were not culturally diffused with other tribes. This distinguished the Setswana linguistic approach from other African languages. The meaning of Setswana months is fast fading and is no longer applicable to their language structure. The lexicographer should keep this in mind in decisions making regarding lemmatization. The following traditional words with their meanings given in brackets are contrasted with their loan forms. Other cultures are being adopted for the following reasons:

- **Industrialization**. Many of the Batswana people moved away from their place of origin to seek employment in industrial areas.
- **Scientific invention.** The Batswana people adopted additional terminology that did not exist in the past and borrowed words to fit into the modern world.

The keywords *ke* (rank 1) with a frequency of (112, 19) and *lo-* (rank 31) frequency of 15,298, are repeated more often in the text than it is expected. The prefix *lo-* is more formal and is used more often in the written than in the spoken texts. Thus *lo-* in the spoken texts signals formal speech, as the subject concord is derived from the prefix of the noun. For example, 'Lo*gong lwa mofiri* **lo** *robega*' (When wood of the mofiri tree breaks up) as compared to 'Legong la mofiri ***le*** robega' (When wood of the mofiri tree breaks up).

Project 9: A traditionalist in Mabopane.

The narrator was born in 1962 in Pelindaba, Pretoria. The narrator's parents were employed as a gardener and a domestic worker and they never stayed with their parents during weekdays. The parents only came home during month-ends. He completed his primary school and secondary education. The narrator understands the Setswana language structures and attaches meaningful interpretation to his words to benefit the audience or readers. He appreciates how a language enriches the life of both the adults and youth of the Batswana people. He identifies certain hidden meanings that could be understood by real Batswana people and therefore has a lot of respect for his language. The narrator reveals the authentic lifestyle of the Batswana people, who based their prosperity on nature and animals. He preserves his environment for his own benefit and shows how the environment contributes to certain lifestyles.

The keywords *ko* (there), *moo* (there), *goo* (there) have been used repeatedly and interchangeable. The narrator has a certain language style that is influenced by the Sekwena dialect where the above words would have been presented in the standard language as *kwa* (there). He frequently uses words like *tlhakalantsuke* (mess), *meraferafe* (nations and nations), *dijarajara* (years and years) as a sign of language enrichment.

Project 10: A particular man in a village.

The narrator does not reveal where comes from nor his age. He only narrates on his brother's life who originates from Kgwadibeng in Hammanskraal. The narrator is both traditional and modern. The language use has no cultural background, but is infiltrated by terms that are linked to the environment in which he lives. The food that he eats is Setswana traditional and influences his typical life style. He mixes languages when addressing his experiences. His language is influenced by the environment in which he lives. He does not visualize any life changes. The language he uses is centred on events, experiences and the environment. He is a Setswana speaker by birth but influenced by the environment in which he lives. The key words *kereya* (to find) appears 11 times in the text and *mara* (but) appears 20 times. These words are loanwords and are more often used than it is expected in the oral text.

Project 11: An educator in a particular school.

The narrator was born in Priska near Makapanstad. He is the fifth child. His parents were neither rich nor poor, but largely dependent on farming and livestock as a means for survival. The narrator had a very good background. He grew up in a village under good conditions, and used pure Setswana because of his educational background. He uses his language with pride and an intense understanding of Setswana dynamics. He has a good command of the Setswana language and is able to motivate youths in his place.

He is committed to his career and uses the language without diverting or switching to other languages. He does not reveal his cultural identity and he does not mention his religion. His language is not influenced by modern life and he uses borrowed words as alternatives to modern technology and scientific inventions. The narrator is living within modern and educational parameters; hence he does not mix languages easily.

The keywords *bone* (seen) (rank 41) with a frequency of 9,891 and *bona* (see) (rank 31) with a frequency of 3,336 is repeated more often in the text than it is to be expected. The pronouns *bone* (they) and *bona* (they) refer to the second and the third person, which does not signal a dialogue. The keyword *gore* (so that) is also used very often to join the narrator's sentences, for example, substantiate or support his statements. Thus these keywords, *bona*, *bone* and *gore* are regarded as positive keys as they are used more often than it is expected in the spoken corpus.

Project 12: A lady principal at a particular school.

The narrator was born at Priska near Makapanstad. She was the third child of five children. She is from a poor family background. Her parents were solely depended on farming as a means of survival. She was well trusted in handling the school finances. Through her dedication and hard work she was promoted to principal. She is a good narrator and remembers most dates and names of events as they arise. She liked her schooling and can face challenges as they arise. She is the type of a person who acclimatizes easily. She is strategic and if things are not in her favour, she seeks for an alternative. She always wants things to happen according to her plan. She is very proud of her language, but willing to communicate with others in their language in order to learn their languages. She is unselfish, but shares and motivates people through her determination. Her perseverance has brought success in her life. She is a teacher by profession and she uses language with great care.

The keyword *ke* (rank 1) frequency of 1.635 occupies the highest rank in the Setswana oral corpus since it refers to the first person singular pronoun and has been repeated more often than it is expected in the text.

Project 13: Praise at a thanks-giving party in Mafikeng

The purpose of the thanks-giving party was to honour and congratulate the celebrity for the role she played in the upbringing of her children as well as her grandchildren. She is regarded as a strong, humble person. The narrator uses the language with surety and does not doubt her language structures. She makes use of borrowed words to make the content appealing, not out of ignorance. Her cultural background can be detected by her phrasing. She does not reveal her religion nor does she align herself with African beliefs. She makes use of borrowed words when she mentions Christian activities. Her language is purely Setswana and she confines herself to Setswana cultural ceremonies and rituals. The narrator does not mention any educational achievements. She is content with her lifestyle and shows no intention of venturing into other areas of existence. The language she uses describes the events and conditions in her life. She is compelled to borrow scientific words.

The keyword *bona* (they) (rank 31) is repeated eight times in one sentence to rhyme with the praise. (See extract below, taken from the text.)

> *"Basimane ba tla bona fa ke se na go bona, ba bona eng ke e se ke bone?*
> *Ba bona eng ke e ise ke bone, ba tla bona fa ke seng go bona*"

The use of exclamations such as *m-m, halala, ee, oo, ee* as keywords is often used in Setswana praise to signal happiness and joy. Thus these words are used more frequently than would be expected in this text.

Table 4 below shows a word list of the top 100 words taken from the Setswana oral corpus with the rankings on the left and the frequencies on the right.

Table 4: Top 100 words from the South African oral Setswana Corpus

Rank	**Word**	**Frequency**
1	KE	1.635
2	A	1.227
3	LE	1.191
4	BA	1.101
5	RE	1.095
6	KA	1.090
7	E	1.061
8	GO	1.025
9	O	834
10	NE	531
11	MO	530
12	YA	514
13	GA	479
14	GORE	455
15	DI	440

Rank	**Word**	**Frequency**
51	ME	66
52	EO	62
53	GONA	61
54	THATA	59
55	ELE	57
56	ILE	57
57	JAAKA	54
58	PELE	54
59	ITSE	52
60	TSAMAYA	52
61	BE	51
62	JALO	50
63	PUO	50
64	TSONA	49
65	ENE	47

16	SE	312
17	WA	285
18	SA	275
19	KWA	235
20	MME	229
21	BO	228
22	FA	222
23	GE	220
24	TSE	214
25	NA	190
26	KO	182
27	NNA	180
28	BONA	170
29	TSA	163
30	JAANONG	147
31	LA	127
32	TENG	126
33	FELA	112
34	TLA	108
35	NTSE	99
36	YO	99
37	BANA	97
38	RA	91
39	NAKO	90
40	NENG	87
41	DIRA	86
42	KGOMO	78
43	YONA	78
44	BATHO	77
45	BUA	77
46	KGWEDI	77
47	KGOTSA	76
48	RONA	76
49	SETSWANA	75
50	MOTHO	67
66	MAINA	46
67	NGWANA	46
68	BONE	45
69	JAANA	45
70	DILO	44
71	MORAGO	44
72	FITLHELA	43
73	GAPE	43
74	LENG	43
75	MONGWE	43
76	BATLA	42
77	GAGWE	42
78	JA	42
79	DITLHAKA	41
80	GAE	41
81	GONE	40
82	MODIMO	40
83	NGWAGA	39
84	DIKGOMO	38
85	TLE	38
86	TSHWANETSE	38
87	EE	37
88	SEKOLO	37
89	TSAYA	37
90	TSENA	37
91	FITLHA	36
92	MONNA	36
93	SENA	36
94	LO	35
95	MARA	35
96	SEKOLONG	35
97	TOTA	35
98	DIANE	33
99	MALOME	33
100	MOTSWANA	33

Keyness analysis of the oral versus the written corpus will now be presented in more detail.

3.7 Identifying and evaluating keyness for the South African oral Setswana corpus

Consider the following extract taken from a female educator, one of the 13 oral Setswana projects:

> *"Go tloga moo mme wa ka ya ba gore o nyalwa ko Makapane kwa Meselane koo e leng gore ke koo saleng ka* ***thoma*** *go golagola* ***kogo*** *e tsena sekolo kogo. Morago ga dingwaga ka mathata ka boela ka mo sefaneng sa ka mo gae sa Tsebe se ke belegeng ka ka sona, jaanong go fitlhila ge ke* ***thoma*** *ke dira mo sekolong sa Marula ka 1980, kege e le gore o na ngwaga oo ke fetola sefane sele sa ka sa Meselane, ke boela mo go sa Tsebe se eleng sa bo* ***ntate*** *wa ka, ka gore mme o ne a nna le mathata ka sefane se le sa Tsebe o nna ke sa mmonang ka matlho a ka, o saleng a tlhokofala ke sale lesea. Janong ke bone ge ke dira tiro ya ka mo sekolong ke bona ke dira sentle le barutabana ba bangwe go sena mathata. Ke bona ke rutile dithuto di le tse ntsi nyana mo sekolong se sa Marula se. ke ile ka bo ke ruta se Afrikaanse dingwaga tse masome a mabedi. English ke e rutile dingwaga nyana tse pedi mo klaseng tsa bo standard 4, dingwaga tsa kgale le bo History ke dir utile bo Geography. Beibele le yona ke e rutile ka ruta le* Setswana. *Jaanong* setswana *ke bona le gona jaanong le santse ke se ruta ko mophatong wa grade 4, mo ke kgonang go bona gore bana ba* setswana ***pilapila*** *ba se tlhaela ka gore ga ba kgone go kwala ditlhaka, le ge o ka ba ruta o ba biletsa ditlhaka o kgona go bona fela gore go na le mathata a tseneletseng mo setswaneng, go na mo baneng ba grade 4, thatathata mo eleng gore go a bonagala gore ba nyaka nako e ntsintsi gore ba rutilwe* setswana *se.*

> *Ke santse ke bala grade 4 yona e o ya standard 2 sa kgale ken e ke ntse ke botsa mme wa ka gore ke batla go nna morutabana, endene mo mabakeng a ka ke dire eng, ke be le mosadi, ke tswele pele ntse ke* ***bereka****. Le ge mosadi a sa bereke, ke nne ke tswele pele. Ke bereka go se na mathata. Jaanong ge ke tla mo tabeng yona e gape ya bana ba palelwang ke ditlhaka, ba ke re ditlhaka di a ba paella, le gore ke lemogile gore ba itse orale that ka gore fa o kwadile mola ditlhaka modichokong, ge o ntse o bua le bona o mo raya o re tlaa I mpale mo, o ipotsa gore ka gore neb a bala e le ba bantsi* ***endene*** *o ipotsa gore daramo sentse ba itse ditlhaka. O tlo bala lefoko le lengwe a supile lefoko le eseng la tlhaka tsego mme a le bala la tlhaka tse di riling. Go o mo raya o re tlhaka tse o di bitsang wa di bona naa, o tla dumela a re e e ka a di bona. O mo ree o re bala gape mme o tla go* ***balla*** *lefoko le lengwe le eseng lona moo. Ka gore o utlwile ba bala mafoko a mantis nyana mo la, ba bala le ena. E na o tshwara fela modumo o go mme o fitlhela e le ditlhaka tseo ga se tsona. Se o ke sona se se ntemogisitseng gore bana ba, ba bathe nako e telele. Ke gore o ka re go ka nna le taba ya gore bana rutiwa ditlhaka mo mephatong e. Se o ke sona se se ntemogisitseng gore bana ba, ba bathe nako e telele*”

The boldfaced words are repeated more often as can be expected from the text. The text is signalled by the dominant personal pronoun *ke* (I) and the possessive pronoun *ka* (with) and the third person plural *ba* (they). A quick review shows that these words are typically associated with the verbs. These words are represented by a pattern of repetition in the above text. The keyness function in WordSmith Tools is an ideal for selecting so-called keywords when comparing a dedicated corpus with a general corpus, for example. This keyness function was used to:

- compare oral and written Setswana so as to be able to study and identify words that are typically used in oral versus written communication

- determine to what extent the written and the oral data differ in Setswana, i.e. are there any differences between the spoken and the written Setswana corpora in terms of words most frequently used and in the registers for example, spoken corpus (more informal), and whether loanwords tend to be used more frequently in oral versus written corpora etc.

To establish keyness, Hoey (1991) and Kintsch and Van Dijk (1978) rely essentially on identifying where there is conceptual repetition. According to them, the conceptual repetition helps identify what the text is all about. They (1996:58) state that:

> "The basic principle is that a word-form which is repeated a lot within the text in question will be more likely to be key in it".

The word keyness is concerned with two aspects of terms such as 'positive keyness' and 'negative keyness' which will be presented below in Tables 5, 6 and 7.

Table 5: Positive keyness versus frequency in the oral corpus

Spoken corpus	**Frequency in oral corpus**	**Keyness**
kereya (find)	28	**262.6**
mara (but)	36	**222.7**
aowa (no)	8	**75.9**
thoma (start)	10	**70.9**
nkebe (maybe)	7	**65.6**
nyaka (want)	7	**65.6**
endene (and then)	6	**56.3**
ntate (father)	25	**54.8**
tjelete (money)	5	**46.9**
feisi (fist)	7	**43.9**

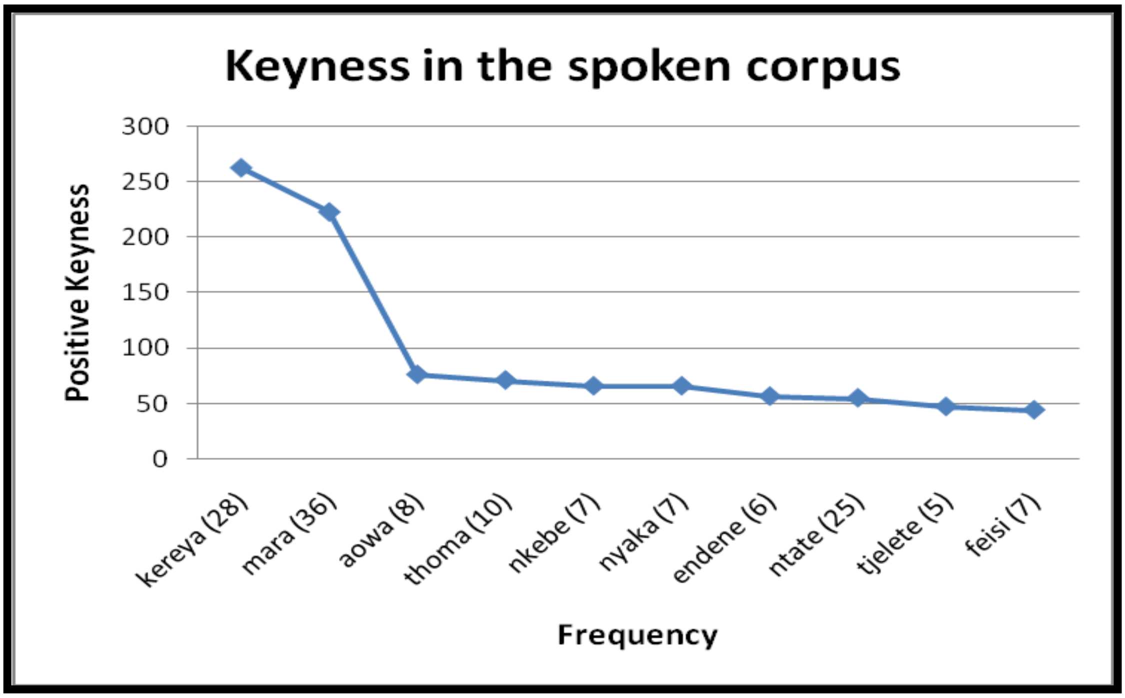


Figure 8: Graphical representation of positive keyness versus frequency in the spoken corpus

Positive keyness refers to those words used much more frequently than it is expected in a given text. The most prominent ones with the highest frequency will firstly be discussed i.e. words like *kereya* (28) and *mara* (36). These words stand out as being unusually frequent and have the highest keyness as shown in Figure 8. Taking the second example of words in the same category like *ntate* (25) and *thoma* (10), one realises the inconsistencies in terms of the likeness between frequency and keyness. One notices that the majority of these words such as *kereya* and *mara* are loanwords. For example, the words *kereya* and *mara* derive from the Afrikaans words '*kry*' and '*maar*' respectively while words such as *ntate* and *thoma* are influences from other dialects and other Sotho languages such as Sepedi and Southern Sotho. These loanwords have the lowest keyness. It is also important to note that these words are used when people are in a more relaxed atmosphere, for example, we often do not find words such as *bona*, *lemororo*, *nyaa* and *simolola* in their regular use in the language. It is important to state that dialects of other regions or social classes be taken into consideration during dictionary compilation. For example, *ntate* is more frequently used than *rre* but both words refer to the word (father) where *ntate* is more often spoken by the Bafokeng in Rustenburg. Conclusion can thus be drawn that there is no clear link between the word frequency and the positive keyness in the oral corpus. The spoken words in Table 5 will now be compared with their more natural counterparts in the language. Consider Table 6 in this regard:

Table 6: Words and their more natural counterparts in the language

Spoken corpus	**Written- corpus**
kereya (find)	*bona*
mara (but)	*Lefa/ le mororo*
aowa (no)	*nyaa*
thoma (start)	*simolola*
nkebe (maybe)	*nkabo*
nyaka (want)	*batla/ eletsa*
endene (and then)	*jaanong*
ntate (father)	*rre*
tjelete (money)	*madi*
feisi (fist)	*lebole*

As with the earlier analysis, negative keyness versus word frequency as indicated in Table 7. Consider the following:

Table 7: Negative keyness versus frequency in the oral corpus

Word	**Frequency in oral corpus**	**Keyness**
gagwe (his/hers)	42	**- 214.9**
jwa (belong to)	35	**- 81.9**
gago (yours)	18	**- 78.1**
jaaka (like)	54	**- 45.6**
jang (how)	10	**- 45.6**

motho (person)	67	**- 39.2**
omongwe (somebody)	46	**- 29.9**
Mopitlwe (March)	6	**- 24**

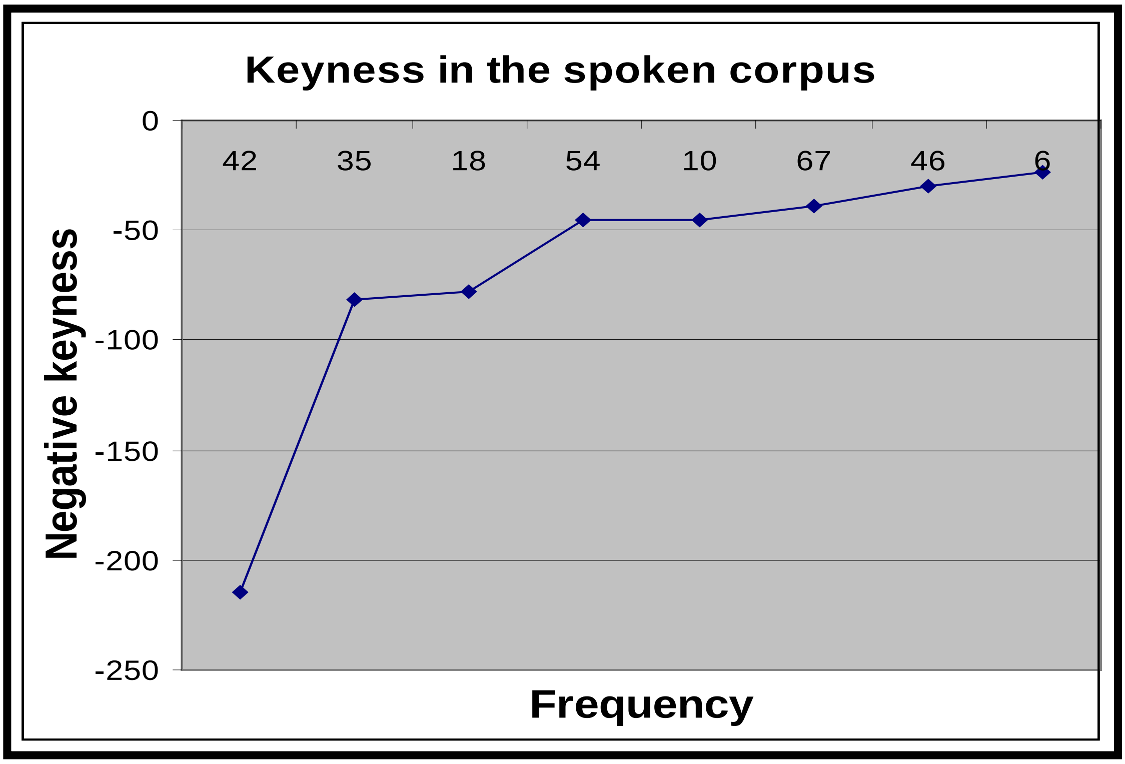


Figure 9: Graphical representation of negative keyness

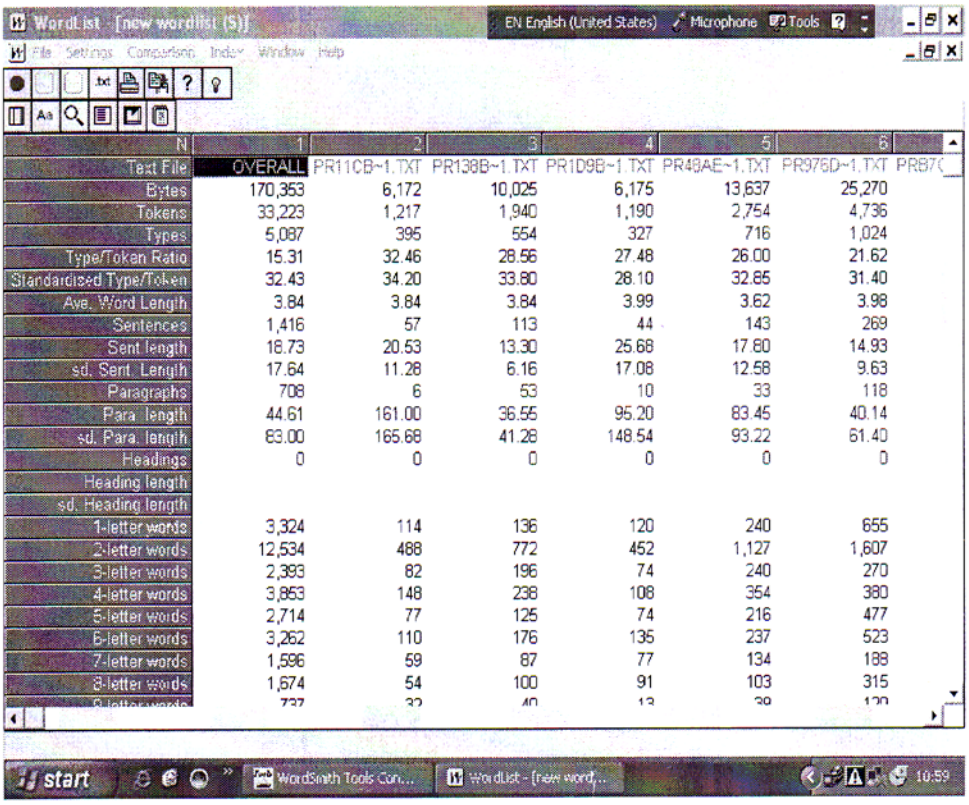
The negative keys are those words used much less frequently than is expected in the oral corpus. Scott (1999) comments that:

> "A word which is negatively key occurs less often than would be expected by chance in comparison with the reference corpus".

One can follow the same approach in Table 7 of comparing word frequency in relation to the Keyness function, taking for example words with the highest frequencies such as *motho* (67), and *gagwe* (42), with keyness of (-39), and (-214.9) respectively. The difference between *motho* and *gagwe* in terms of their keyness does not show any correlation between keyness and frequency. If one goes further in comparing for example, words such as *Mopitlwe* (6) with the lowest frequency and *motho* (67) with the highest frequency, it is important to state that *Mopitlwe* with the lowest frequency has the highest keyness while *motho* with the highest frequency has a relatively high keyness. Thus it is important to conclude from the above given examples that there is no correlation between the negative keyness and the word frequency in the spoken corpus.

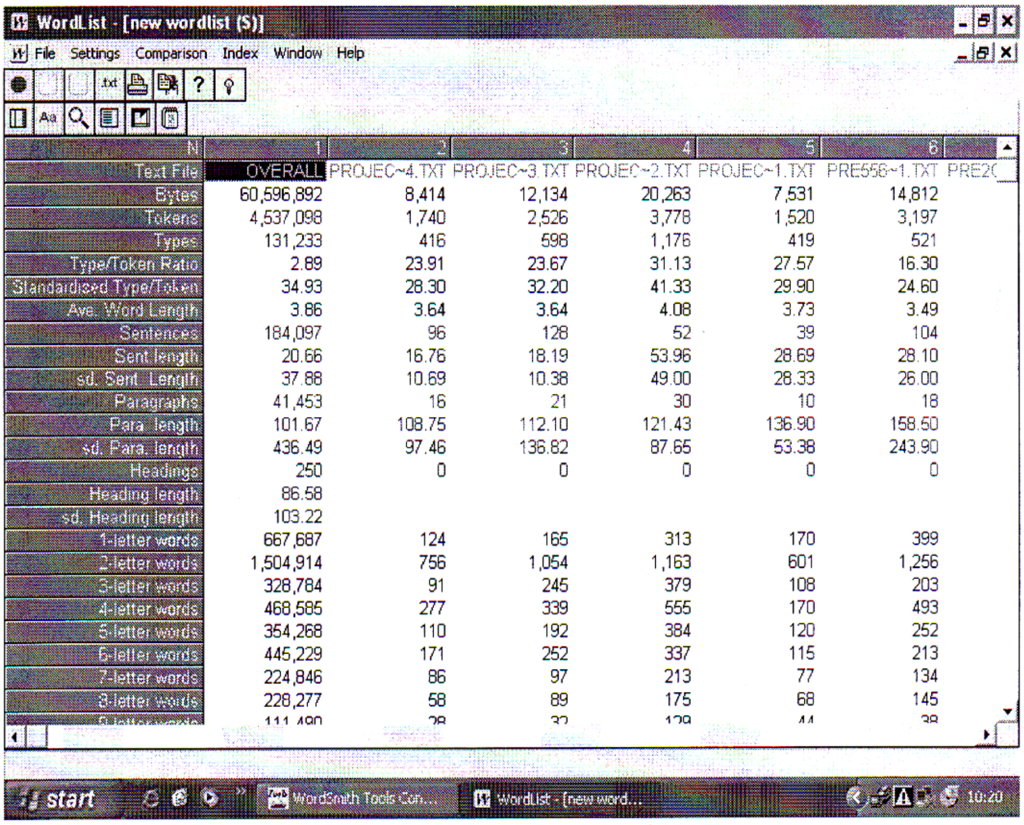
3.8 Comparison between the South African oral corpus and the South African written corpus

The aim here is to develop accounts of the variation between the South African oral and the written Setswana corpora. We will do this by distinguishing between the oral and the written corpora on the base of the ranks of their top 100 tokens. The study is done firstly by giving the statistical analysis of both the oral and the written corpora and then the word lists for the top 100 tokens from each corpus are compared using WordSmith Tools. Consider the following statistical information below:



N	1	2	3	4	5	6	
Text File	OVERALL	PR11CB~1.TXT	PR138B~1.TXT	PR1D9B~1.TXT	PR48AE~1.TXT	PR976D~1.TXT	PRB7(
Bytes	170,353	6,172	10,025	6,175	13,637	25,270	
Tokens	33,223	1,217	1,940	1,190	2,754	4,736	
Types	5,087	395	554	327	716	1,024	
Type/Token Ratio	15.31	32.46	28.56	27.48	26.00	21.62	
Standardised Type/Token	32.43	34.20	33.80	28.10	32.85	31.40	
Ave. Word Length	3.84	3.84	3.84	3.99	3.62	3.98	
Sentences	1,416	57	113	44	143	269	
Sent length	18.73	20.53	13.30	25.68	17.80	14.93	
sd. Sent. Length	17.64	11.28	6.16	17.08	12.58	9.63	
Paragraphs	708	6	53	10	33	118	
Para. length	44.61	161.00	36.55	95.20	83.45	40.14	
sd. Para. length	83.00	165.68	41.28	148.54	93.22	61.40	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	3,324	114	136	120	240	655	
2-letter words	12,534	488	772	452	1,127	1,607	
3-letter words	2,393	82	196	74	240	270	
4-letter words	3,853	148	238	108	354	380	
5-letter words	2,714	77	125	74	216	477	
6-letter words	3,262	110	176	135	237	523	
7-letter words	1,596	59	87	77	134	188	
8-letter words	1,674	54	100	91	103	315	
9-letter words	737	32	40	13	39	120	

Figure 10: Statistical analysis of the Setswana oral corpus in WordSmith Tools



WordList - [new wordlist (S)]
File Settings Comparison Index Window Help
.txt
?
start
WordSmith Tools Con...
WordList - [new word...
10:20

N	1	2	3	4	5	6	
Text File	OVERALL	PROJEC~4.TXT	PROJEC~3.TXT	PROJEC~2.TXT	PROJEC~1.TXT	PRE556~1.TXT	PRE2(
Bytes	60,596,892	8,414	12,134	20,263	7,531	14,812	
Tokens	4,537,098	1,740	2,526	3,778	1,520	3,197	
Types	131,233	416	598	1,176	419	521	
Type/Token Ratio	2.89	23.91	23.67	31.13	27.57	16.30	
Standardised Type/Token	34.93	28.30	32.20	41.33	29.90	24.60	
Ave. Word Length	3.86	3.64	3.64	4.08	3.73	3.49	
Sentences	184,097	96	128	52	39	104	
Sent length	20.66	16.76	18.19	53.96	28.69	28.10	
sd. Sent. Length	37.88	10.69	10.38	49.00	28.33	26.00	
Paragraphs	41,453	16	21	30	10	18	
Para. length	101.67	108.75	112.10	121.43	136.90	158.50	
sd. Para. length	436.49	97.46	136.82	87.65	53.38	243.90	
Headings	250	0	0	0	0	0	
Heading length	86.58						
sd. Heading length	103.22						
1-letter words	667,687	124	165	313	170	399	
2-letter words	1,504,914	756	1,054	1,163	601	1,256	
3-letter words	328,784	91	245	379	108	203	
4-letter words	468,585	277	339	555	170	493	
5-letter words	354,268	110	192	384	120	252	
6-letter words	445,229	171	252	337	115	213	
7-letter words	224,846	86	97	213	77	134	
8-letter words	228,277	58	89	175	68	145	
9-letter words	111,490	28	32	129	44	38	

Figure 11: Statistical analysis of the Setswana written corpus in WordSmith Tools

In the following section, we will give an account of the contrast of the top 100 items between the South African oral corpus, South African written corpus and the Botswana text corpora. The word lists in Tables 8, 9, 10, 11, 12 and 13 are ordered by rank that is the top items are the most frequent used words.

In Table 8, for example the top 100 words of the South African oral corpus are given in boldface compared to their ranks in the South African written corpus which is sorted in ascending order of ranks.

Table 8: Comparison of the top 100 items between the South African oral and the South African written corpus

South African oral corpus		South African written corpus	
Rank	Word	Rank	Word
2	**A**	1	A
8	**GO**	2	GO
3	**LE**	3	LE
7	**E**	4	E
9	**O**	5	O
4	**BA**	6	BA
6	**KA**	7	KA
1	**KE**	8	KE
12	**YA**	9	YA
11	**MO**	10	MO
13	**GA**	11	GA
22	**FA**	12	FA
5	**RE**	13	RE
16	**SE**	14	SE
10	**NE**	15	NE
15	**DI**	16	DI
17	**WA**	17	WA
14	**GORE**	18	GORE
18	**SA**	19	SA
29	**TSA**	20	TSA
19	**KWA**	21	KWA
34	**TLA**	22	TLA
20	**MME**	23	MME
24	**TSE**	24	TSE
21	**BO**	25	BO
31	**LA**	26	LA
77	**GAGWE**	27	GAGWE
36	**YO**	28	YO
27	**NNA**	29	NNA
28	**BONA**	30	BONA
94	**LO**	31	LO
33	**FELA**	32	FELA
25	**NA**	33	NA
35	**NTSE**	34	NTSE
59	**ITSE**	35	ITSE
57	**JAAKA**	36	JAAKA
40	**NENG**	38	NENG
50	**MOTHO**	39	MOTHO
51	**ME**	40	ME
68	**BONE**	41	BONE
44	**BATHO**	42	BATHO
75	**MONGWE**	43	MONGWE
32	**TENG**	45	TENG
45	**BUA**	46	BUA
58	**PELE**	47	PELE
65	**ENE**	49	ENE
41	**DIRA**	50	DIRA

62	**JALO**	52	JALO
71	**MORAGO**	53	MORAGO
54	**THATA**	56	THATA
92	**MONNA**	57	MONNA
48	**RONA**	63	RONA
43	**YONA**	66	YONA
67	**NGWANA**	68	NGWANA
76	**BATLA**	69	BATLA
90	**TSENA**	70	TSENA
39	**NAKO**	71	NAKO
73	**GAPE**	72	GAPE
47	**KGOTSA**	73	KGOTSA
37	**BANA**	75	BANA
30	**JAANONG**	77	JAANONG
89	**TSAYA**	78	TSAYA
60	**TSAMAYA**	79	TSAMAYA
85	**TLE**	80	TLE
70	**DILO**	86	DILO
61	**BE**	93	BE
82	**MODIMO**	95	MODIMO
69	**JAANA**	96	JAANA
74	**LENG**	***102***	***LENG***
86	**TSHWANETSE**	***103***	***TSHWANETSE***
91	**FITLHA**	***108***	***FITLHA***
78	**JA**	***109***	***JA***
64	**TSONA**	***113***	***TSONA***
97	**TOTA**	***116***	***TOTA***
93	**SENA**	***117***	***SENA***
53	**GONA**	***123***	***GONA***
81	**GONE**	***125***	***GONE***
80	**GAE**	***129***	***GAE***
72	**FITLHELA**	***130***	***FITLHELA***
52	**EO**	***133***	***EO***
38	**RA**	***138***	***RA***
56	**ILE**	***157***	***ILE***
42	**KGOMO**	***163***	***KGOMO***
63	**PUO**	***174***	***PUO***
84	**DIKGOMO**	***191***	***DIKGOMO***
87	**EE**	***202***	***EE***
88	**SEKOLO**	***267***	***SEKOLO***
83	**NGWAGA**	***279***	***NGWAGA***
66	**MAINA**	***307***	***MAINA***
26	**KO**	***350***	***KO***
96	**SEKOLONG**	***366***	***SEKOLONG***
49	**SETSWANA**	***369***	***SETSWANA***
46	**KGWEDI**	***380***	***KGWEDI***
55	**ELE**	***475***	***ELE***
98	**DIANE**	***485***	***DIANE***
99	**MALOME**	***739***	***MALOME***
23	**GE**	***834***	***GE***
79	**DITLHAKA**	***896***	***DITLHAKA***
100	**MOTSWANA**	***1317***	***MOTSWANA***
95	**MARA**	***5540***	***MARA***

Studying the number of items which occur versus the number of ousted items in comparing the top 100 items in the South African oral corpus and the South African written corpus, it is important to note that 68% of the items in the written corpus are retained while 32% of the items fall outside the top 100 items in the South African oral corpus, cf. counts 102, 103, 108,….5, 540 for the 32 words *leng*, *tshwanetse*, …. *mara*. If we consider ousted items in the South African written corpus in Table 8, we conclude that 30 of the 32 ousted items are still very high falling in the range 101-896 while two of the items *Motswana* and especially *mara* rank much lower in the written corpus outside the top 100 items. It is important to note that a word like *mara* is a loanword and is more frequently used in the oral corpus than in the written corpus and is more frequently used where Afrikaans in South Africa is predominantly spoken.

Table 9: Comparison of the top 100 items between the South African written corpus and the South African oral corpus

South African written corpus v/s **South African oral corpus**

Rank	Word	Rank	Word
8	KE	1	KE
1	A	2	A
3	LE	3	LE
6	BA	4	BA
13	RE	5	RE
7	KA	6	KA
4	E	7	E
2	GO	8	GO
5	O	9	O
15	NE	10	NE
10	MO	11	MO
9	YA	12	YA
11	GA	13	GA
18	GORE	14	GORE
16	DI	15	DI
14	SE	16	SE
17	WA	17	WA
19	SA	18	SA
21	KWA	19	KWA
23	MME	20	MME
25	BO	21	BO
12	FA	22	FA
24	TSE	24	TSE

33	NA	25	NA
29	NNA	27	NNA
30	BONA	28	BONA
20	TSA	29	TSA
77	JAANONG	30	JAANONG
26	LA	31	LA
45	TENG	32	TENG
32	FELA	33	FELA
22	TLA	34	TLA
34	NTSE	35	NTSE
28	YO	36	YO
75	BANA	37	BANA
71	NAKO	39	NAKO
38	NENG	40	NENG
50	DIRA	41	DIRA
66	YONA	43	YONA
42	BATHO	44	BATHO
46	BUA	45	BUA
73	KGOTSA	47	KGOTSA
63	RONA	48	RONA
39	MOTHO	50	MOTHO
40	ME	51	ME
56	THATA	54	THATA
36	JAAKA	57	JAAKA
47	PELE	58	PELE
35	ITSE	59	ITSE
79	TSAMAYA	60	TSAMAYA
93	BE	61	BE
52	JALO	62	JALO
49	ENE	65	ENE
68	NGWANA	67	NGWANA
41	BONE	68	BONE
96	JAANA	69	JAANA
86	DILO	70	DILO
53	MORAGO	71	MORAGO
72	GAPE	73	GAPE
43	MONGWE	75	MONGWE
69	BATLA	76	BATLA
27	GAGWE	77	GAGWE
95	MODIMO	82	MODIMO
80	TLE	85	TLE
78	TSAYA	89	TSAYA
70	TSENA	90	TSENA
57	MONNA	92	MONNA
31	LO	94	LO
58	ENG	*101*	*ENG*
37	JWA	*102*	*JWA*
65	NTLHA	*104*	*NTLHA*
97	NNGWE	*109*	*NNGWE*
92	RAYA	*110*	*RAYA*
90	RATA	*112*	*RATA*
84	BILE	*130*	*BILE*
82	ENA	*131*	*ENA*
74	LETSATSI	*135*	*LETSATSI*
98	JO	*142*	*JO*

81	**MAFOKO**	*145*	*MAFOKO*
76	**MOSADI**	*146*	*MOSADI*
83	**SENGWE**	*152*	*SENGWE*
85	**SENTLE**	*153*	*SENTLE*
100	**TSOTLHE**	*154*	*TSOTLHE*
48	**TSWA**	*160*	*TSWA*
60	**SETSE**	*166*	*SETSE*
89	**KANA**	*170*	*KANA*
94	**KAE**	*180*	*KAE*
44	**GAGO**	*190*	*GAGO*
59	**UTLWA**	*198*	*UTLWA*
62	**WENA**	*221*	*WENA*
54	**GONNE**	*253*	*GONNE*
64	**RILE**	*259*	*RILE*
51	**KGOSI**	*281*	*KGOSI*
88	**GODIMO**	*342*	*GODIMO*
87	**JANG**	*347*	*JANG*
55	**I**	*436*	*I*
67	**PELO**	*453*	*PELO*
91	**MATLHO**	*512*	*MATLHO*
99	**IWA**	*3055*	*IWA*

If one follows the same approach in Table 9 of comparing the top 100 items in the South African written corpus and the South African oral corpus, one concludes that 69% (one item was not considered) of the South African oral corpus is retained while 31% of the items are thrown outside the top 100 items of the South African written corpus. What is important to note is that 30 of the 31 ousted items are still very high falling in the range 101-512 while the item like *iwa* falling slightly higher. It is important to note that the social environment in which we find ourselves influences the way we use the language. For example, being in a heterogeneous language, people turn to code-switch from one language to another. Consider the following examples below:

- Heart problem instead of *matlhoko a pelo*
- Eye-lids instead of *ditlhaka tsa matlho*

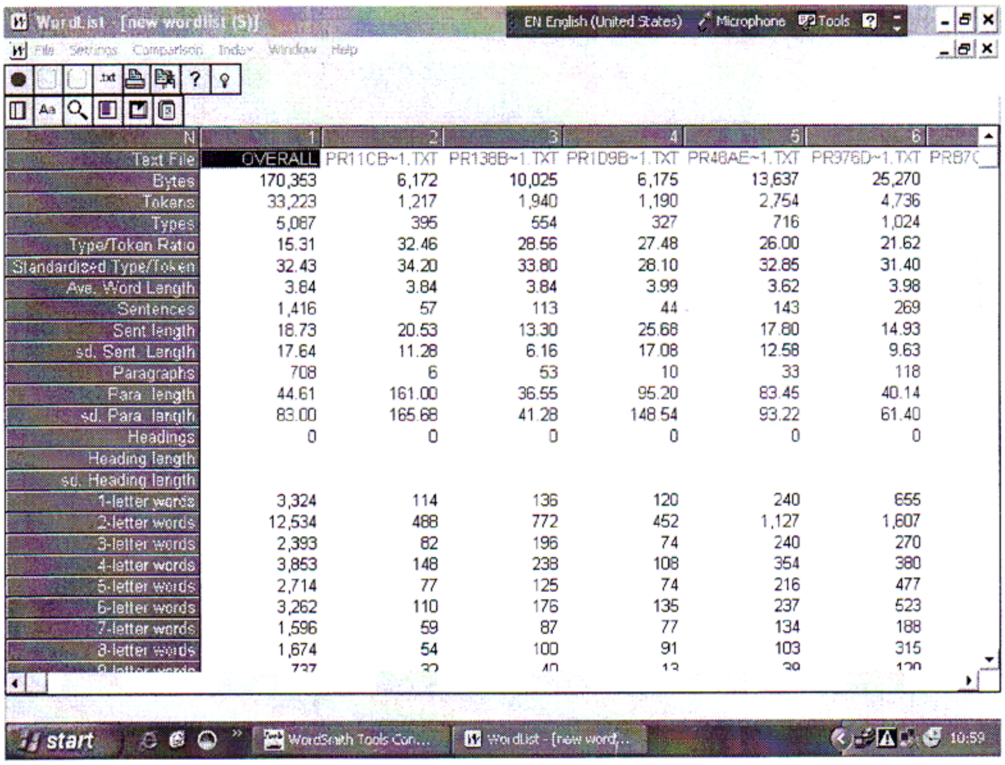
The words *pelo* and *matlho* are more formal and are more often used in the written corpus than in the spoken corpus. It is important to note that people in conversations turn to pronounce words such as *iwa* as *yiwa*. Consider the following example:

- *Go* ***yiwa*** *gae* instead of *go* ***iwa*** *gae* (They went home)[1]

This suggests that the item *iwa* is more frequently used in the written corpus than in the oral corpus.

3.9 Comparison between the Botswana text corpus and the South African oral corpus

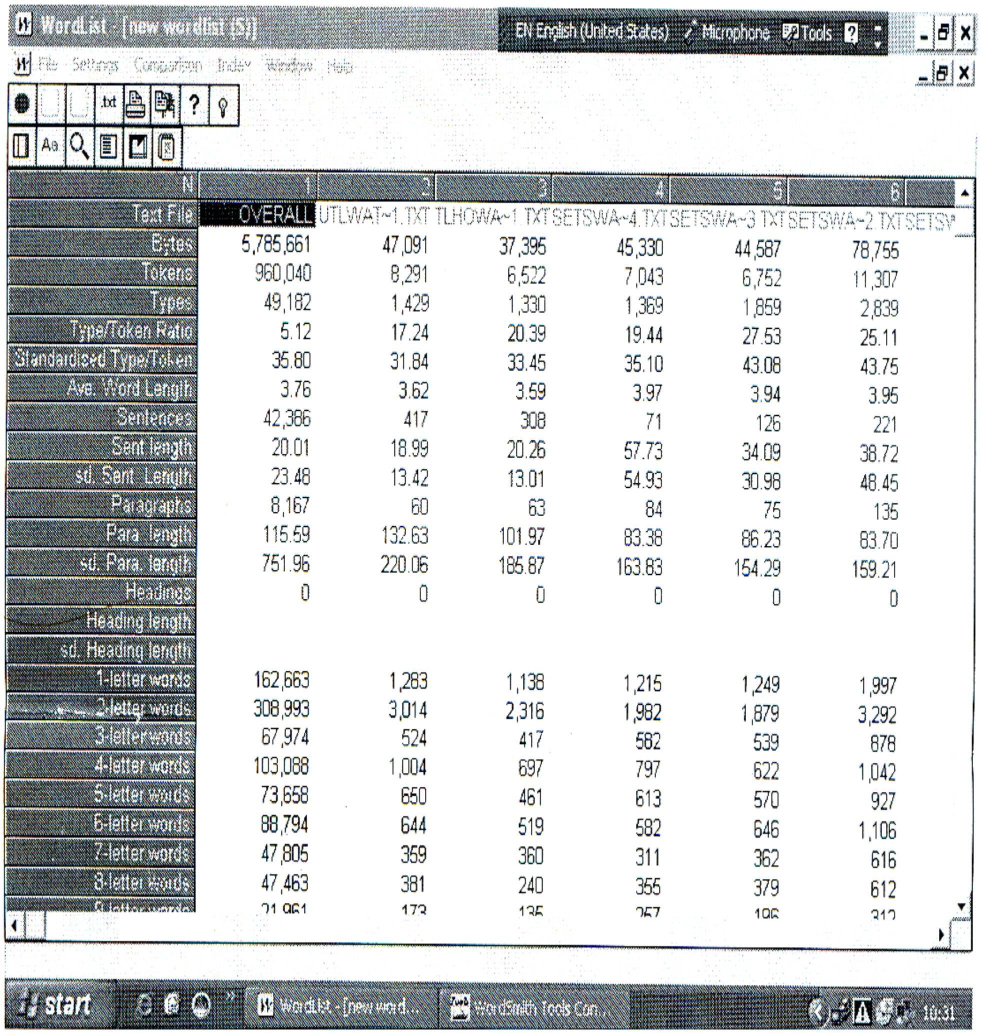
Figure 12 represents the statistical analysis of the South African oral corpus in WordSmith Tool consisting of 33,223 tokens.



N	1	2	3	4	5	6	
Text File	OVERALL	PR11CB~1.TXT	PR138B~1.TXT	PR1D9B~1.TXT	PR48AE~1.TXT	PR376D~1.TXT	PR87C
Bytes	170,353	6,172	10,025	6,175	13,637	25,270	
Tokens	33,223	1,217	1,940	1,190	2,754	4,736	
Types	5,087	395	554	327	716	1,024	
Type/Token Ratio	15.31	32.46	28.56	27.48	26.00	21.62	
Standardised Type/Token	32.43	34.20	33.80	28.10	32.85	31.40	
Ave. Word Length	3.84	3.84	3.84	3.99	3.62	3.98	
Sentences	1,416	57	113	44	143	269	
Sent length	18.73	20.53	13.30	25.68	17.80	14.93	
sd. Sent. Length	17.64	11.28	6.16	17.08	12.58	9.63	
Paragraphs	708	6	53	10	33	118	
Para. length	44.61	161.00	36.55	95.20	83.45	40.14	
sd. Para. length	83.00	165.68	41.28	148.54	93.22	61.40	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	3,324	114	136	120	240	655	
2-letter words	12,534	488	772	452	1,127	1,607	
3-letter words	2,393	82	196	74	240	270	
4-letter words	3,853	148	238	108	354	380	
5-letter words	2,714	77	125	74	216	477	
6-letter words	3,262	110	176	135	237	523	
7-letter words	1,596	59	87	77	134	188	
8-letter words	1,674	54	100	91	103	315	
9-letter words	737	32	40	13	39	120	

Figure 12: Statistical analysis of the Setswana oral corpus in WordSmith Tools

Figure 13 below represents the statistical analysis of the Botswana text corpus in WordSmith Tools consisting of 960,040 tokens.



WordList - [new wordlist (5)]
EN English (United States)
Microphone
Tools
File Settings Comparison Index Window Help
start
WordList - [new word...
WordSmith Tools Con...
10:31

N	1	2	3	4	5	6
Text File	OVERALL	UTLWAT~1.TXT	TLHOWA~1.TXT	SETSWA~4.TXT	SETSWA~3.TXT	SETSWA~2.TXT
Bytes	5,785,661	47,091	37,395	45,330	44,587	78,755
Tokens	960,040	8,291	6,522	7,043	6,752	11,307
Types	49,182	1,429	1,330	1,369	1,859	2,839
Type/Token Ratio	5.12	17.24	20.39	19.44	27.53	25.11
Standardised Type/Token	35.80	31.84	33.45	35.10	43.08	43.75
Ave. Word Length	3.76	3.62	3.59	3.97	3.94	3.95
Sentences	42,386	417	308	71	126	221
Sent length	20.01	18.99	20.26	57.73	34.09	38.72
sd. Sent. Length	23.48	13.42	13.01	54.93	30.98	48.45
Paragraphs	8,167	60	63	84	75	135
Para. length	115.59	132.63	101.97	83.38	86.23	83.70
sd. Para. length	751.96	220.06	185.87	163.83	154.29	159.21
Headings	0	0	0	0	0	0
Heading length						
sd. Heading length						
1-letter words	162,663	1,283	1,138	1,215	1,249	1,997
2-letter words	308,993	3,014	2,316	1,982	1,879	3,292
3-letter words	67,974	524	417	582	539	878
4-letter words	103,088	1,004	697	797	622	1,042
5-letter words	73,658	650	461	613	570	927
6-letter words	88,794	644	519	582	646	1,106
7-letter words	47,805	359	360	311	362	616
8-letter words	47,463	381	240	355	379	612
9-letter words	21,961	173	135	257	196	312

Figure 13: Statistical analysis of the Botswana text corpus in WordSmith Tools

When the South African oral corpus and the Botswana text corpora in Figures 12 and 13 are compared in terms of the number of tokens, the Botswana text corpus is 28, 9 times larger than the South African oral corpus. Compare the following:

Table 10: Comparison of the top 100 items between the Botswana text corpus and the South African oral corpus

Botswana corpus v/s South African oral corpus

Rank	Word	Rank	Word
8	**KE**	1	KE

1	**A**	2	A
3	**LE**	3	LE
6	**BA**	4	BA
14	**RE**	5	RE
7	**KA**	6	KA
4	**E**	7	E
2	**GO**	8	GO
5	**O**	9	O
12	**NE**	10	NE
9	**MO**	11	MO
10	**YA**	12	YA
11	**GA**	13	GA
17	**GORE**	14	GORE
16	**DI**	15	DI
15	**SE**	16	SE
19	**WA**	17	WA
18	**SA**	18	SA
20	**KWA**	19	KWA
24	**MME**	20	MME
21	**BO**	21	BO
13	**FA**	22	FA
23	**TSE**	24	TSE
33	**NA**	25	NA
27	**NNA**	27	NNA
31	**BONA**	28	BONA
22	**TSA**	29	TSA
94	**JAANONG**	30	JAANONG
28	**LA**	31	LA
48	**TENG**	32	TENG
29	**FELA**	33	FELA
25	**TLA**	34	TLA
35	**NTSE**	35	NTSE
30	**YO**	36	YO
42	**BANA**	37	BANA
63	**NAKO**	39	NAKO
50	**NENG**	40	NENG
65	**DIRA**	41	DIRA
39	**BATHO**	44	BATHO
53	**BUA**	45	BUA
85	**KGOTSA**	47	KGOTSA
92	**RONA**	48	RONA
37	**MOTHO**	50	MOTHO
78	**ME**	51	ME
60	**THATA**	54	THATA
43	**JAAKA**	57	JAAKA
73	**PELE**	58	PELE
34	**ITSE**	59	ITSE
81	**TSAMAYA**	60	TSAMAYA
58	**JALO**	62	JALO
38	**ENE**	65	ENE
46	**NGWANA**	67	NGWANA
36	**BONE**	68	BONE
66	**MORAGO**	71	MORAGO
89	**GAPE**	73	GAPE
49	**MONGWE**	75	MONGWE

67	**BATLA**	76	BATLA
32	**GAGWE**	77	GAGWE
76	**JA**	78	JA
86	**GONE**	81	GONE
82	**TLE**	85	TLE
74	**TSAYA**	89	TSAYA
70	**TSENA**	90	TSENA
52	**MONNA**	92	MONNA
44	**LO**	94	LO
93	**TOTA**	97	TOTA
47	**ENG**	***101***	***ENG***
51	**JWA**	***102***	***JWA***
80	**RAYA**	***110***	***RAYA***
96	**BILE**	***130***	***BILE***
72	**LETSATSI**	***135***	***LETSATSI***
95	**TIRO**	***138***	***TIRO***
69	**MOSADI**	***146***	***MOSADI***
83	**SENGWE**	***152***	***SENGWE***
84	**SENTLE**	***153***	***SENTLE***
45	**TSWA**	***160***	***TSWA***
64	**SETSE**	***166***	***SETSE***
55	**KANA**	***170***	***KANA***
91	**KAE**	***180***	***KAE***
77	**GAGO**	***190***	***GAGO***
100	**MADI**	***192***	***MADI***
79	**UTLWA**	***198***	***UTLWA***
87	**WENA**	***221***	***WENA***
56	**RILE**	***259***	***RILE***
97	**YONE**	***267***	***YONE***
90	**B**	***269***	***B***
59	**MMA**	***284***	***MMA***
62	**JANG**	***347***	***JANG***
26	**I**	***436***	***I***
88	**PELO**	***453***	***PELO***
75	**R**	***528***	***R***
40	**L**	***756***	***L***
57	**T**	***845***	***T***
61	**TWE**	***870***	***TWE***
71	**BAITHUTI**	***887***	***BAITHUTI***
99	**THUSA**	***1156***	***THUSA***
68	**C**	***2321***	***C***
54	**TLAA**	***4889***	***TLAA***

If one compares the top 100 items in the Botswana text corpus and the South African oral corpus, it is important to observe that 68% of the items in the South African oral corpus are retained while 32% of the items are thrown outside the top 100 items in the Botswana text corpus. If one considers the ousted items in Table 10, one concludes that 29 of the 32 items are still very high falling in the range 101-861 while 2 of these ousted items like *thusa* and *tlaa* falling too far outside the top 100 items. It is important to note that the Botswana people use more spoken variants than the written ones. For example, they pronounce the future morpheme *tla* as *tlaa* (will) and *e tla* as *tlaa* (come) which is in line with the revised Setswana spelling rules while the South African people use the future morpheme *tla* more frequently in both the spoken and the written corpus. Thus the item *tlaa* is more frequently used in the Botswana corpus and very seldom in the South African written corpus. Given the information above, it is important to conclude that the top 100 items in the Botswana text corpus and the South African oral corpus differ substantially.

Table 11: Comparison of the top 100 items between the South African oral corpus and the Botswana text corpora

South African oral corpus v/s Botswana corpus

Rank	**Word**	Rank	Word
2	**A**	1	A
8	**GO**	2	GO
3	**LE**	3	LE
7	**E**	4	E
9	**O**	5	O
4	**BA**	6	BA
6	**KA**	7	KA
1	**KE**	8	KE
11	**MO**	9	MO
12	**YA**	10	YA
13	**GA**	11	GA
10	**NE**	12	NE
22	**FA**	13	FA
5	**RE**	14	RE
16	**SE**	15	SE
15	**DI**	16	DI
14	**GORE**	17	GORE
18	**SA**	18	SA
17	**WA**	19	WA
19	**KWA**	20	KWA
21	**BO**	21	BO
29	**TSA**	22	TSA
24	**TSE**	23	TSE
20	**MME**	24	MME
34	**TLA**	25	TLA
27	**NNA**	27	NNA
31	**LA**	28	LA
33	**FELA**	29	FELA
36	**YO**	30	YO
28	**BONA**	31	BONA
77	**GAGWE**	32	GAGWE

25	**NA**	33	NA
59	**ITSE**	34	ITSE
35	**NTSE**	35	NTSE
68	**BONE**	36	BONE
50	**MOTHO**	37	MOTHO
65	**ENE**	38	ENE
44	**BATHO**	39	BATHO
37	**BANA**	42	BANA
57	**JAAKA**	43	JAAKA
94	**LO**	44	LO
67	**NGWANA**	46	NGWANA
32	**TENG**	48	TENG
75	**MONGWE**	49	MONGWE
40	**NENG**	50	NENG
92	**MONNA**	52	MONNA
45	**BUA**	53	BUA
62	**JALO**	58	JALO
54	**THATA**	60	THATA
39	**NAKO**	63	NAKO
41	**DIRA**	65	DIRA
71	**MORAGO**	66	MORAGO
76	**BATLA**	67	BATLA
90	**TSENA**	70	TSENA
58	**PELE**	73	PELE
89	**TSAYA**	74	TSAYA
78	**JA**	76	JA
51	**ME**	78	ME
60	**TSAMAYA**	81	TSAMAYA
85	**TLE**	82	TLE
47	**KGOTSA**	85	KGOTSA
81	**GONE**	86	GONE
73	**GAPE**	89	GAPE
48	**RONA**	92	RONA
97	**TOTA**	93	TOTA
30	**JAANONG**	94	JAANONG
70	DILO	***101***	***DILO***
69	JAANA	***107***	***JAANA***
72	FITLHELA	***110***	***FITLHELA***
86	TSHWANETSE	***118***	***TSHWANETSE***
56	ILE	***126***	***ILE***
93	SENA	***134***	***SENA***
74	LENG	***137***	***LENG***
38	RA	***147***	***RA***
63	PUO	***150***	***PUO***
80	GAE	***154***	***GAE***
52	EO	***173***	***EO***
88	SEKOLO	***177***	***SEKOLO***
87	EE	***181***	***EE***
96	SEKOLONG	***191***	***SEKOLONG***
91	FITLHA	***198***	***FITLHA***
42	KGOMO	***205***	***KGOMO***
43	YONA	***212***	***YONA***
66	MAINA	***225***	***MAINA***
64	TSONA	***236***	***TSONA***
84	DIKGOMO	***266***	***DIKGOMO***

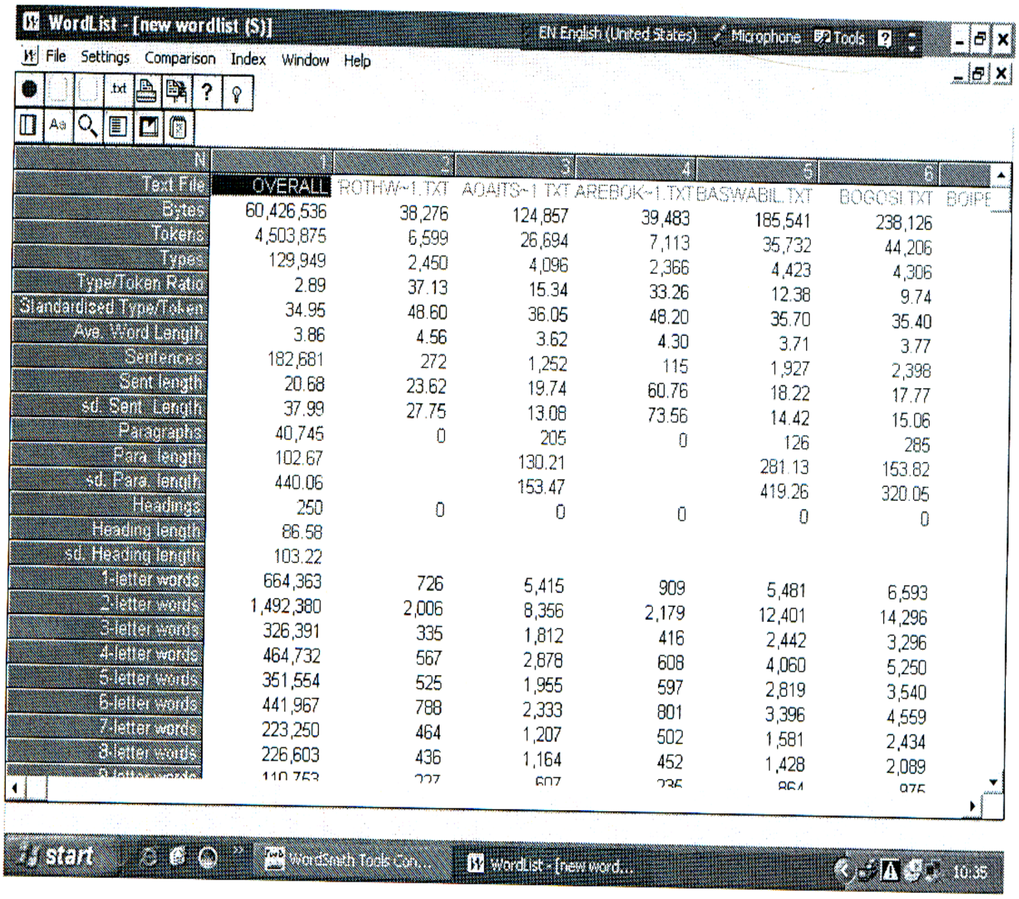
82	MODIMO	*278*	*MODIMO*
49	SETSWANA	*283*	*SETSWANA*
98	DIANE	*333*	*DIANE*
55	ELE	*366*	*ELE*
83	NGWAGA	*399*	*NGWAGA*
61	BE	*405*	*BE*
46	KGWEDI	*420*	*KGWEDI*
26	KO	*452*	*KO*
53	GONA	*489*	*GONA*
99	MALOME	*688*	*MALOME*
100	MOTSWANA	*861*	*MOTSWANA*
79	DITLHAKA	*1249*	*DITLHAKA*
23	GE	*2666*	*GE*
95	MARA	*13072*	*MARA*

If one follows the same approach in Table 11 of comparing the top 100 items in the South African oral corpus and the Botswana text corpus, it is important to note that 66% (two items were not considered) of the items in the Botswana text corpus are retained while 34% of the items are thrown outside the top 100 items in the South African oral corpus. If one considers the ousted items in Table 11, 31 of the 34 items fall very close in the range 101-861 while 3 of the ousted items like *ditlhaka*, *ge* and especially *mara* falling too far the top 100 items in the South African oral corpus. The word *mara* is a loanword (as discussed previously in this section) and is more frequently used in the South African spoken corpus than the written corpus. Thus the word *mara* is less frequently used in the Botswana text corpus.

3.10 Comparison between the South African written corpus and the Botswana text corpus

The study shows how the statistical analysis is able to give insights into how the South African written corpus and the Botswana text corpora differ in terms of their ranks, then the South African and the Botswana word lists generated through the WordSmith Tools will be compared in terms of their ranking orders as illustrated below in Tables 12 and 13.

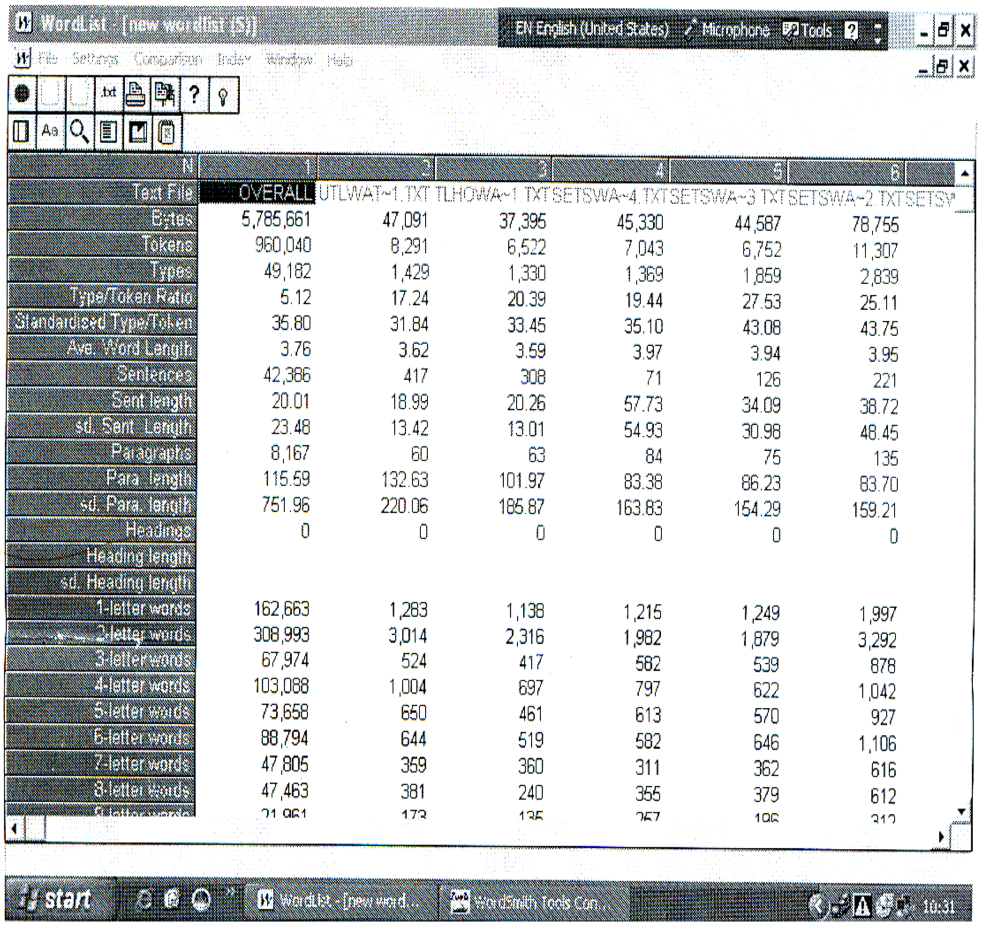
Figure 14 represents the statistical analysis of the South African text corpus in WordSmith Tools comprising of 4,503,875 tokens. Consider the following in this regard:



N	1	2	3	4	5	6	
Text File	OVERALL	ROTHW~1.TXT	AOAITS~1.TXT	AREBOK~1.TXT	BASWABIL.TXT	BOGOSI.TXT	BOIPE
Bytes	60,426,536	38,276	124,857	39,483	185,541	238,126	
Tokens	4,503,875	6,599	26,694	7,113	35,732	44,206	
Types	129,949	2,450	4,096	2,366	4,423	4,306	
Type/Token Ratio	2.89	37.13	15.34	33.26	12.38	9.74	
Standardised Type/Token	34.95	48.60	36.05	48.20	35.70	35.40	
Ave. Word Length	3.86	4.56	3.62	4.30	3.71	3.77	
Sentences	182,681	272	1,252	115	1,927	2,398	
Sent length	20.68	23.62	19.74	60.76	18.22	17.77	
sd. Sent. Length	37.99	27.75	13.08	73.56	14.42	15.06	
Paragraphs	40,745	0	205	0	126	285	
Para. length	102.67		130.21		281.13	153.82	
sd. Para. length	440.06		153.47		419.26	320.05	
Headings	250	0	0	0	0	0	
Heading length	86.58						
sd. Heading length	103.22						
1-letter words	664,363	726	5,415	909	5,481	6,593	
2-letter words	1,492,380	2,006	8,356	2,179	12,401	14,296	
3-letter words	326,391	335	1,812	416	2,442	3,296	
4-letter words	464,732	567	2,878	608	4,060	5,250	
5-letter words	351,554	525	1,955	597	2,819	3,540	
6-letter words	441,967	788	2,333	801	3,396	4,559	
7-letter words	223,250	464	1,207	502	1,581	2,434	
8-letter words	226,603	436	1,164	452	1,428	2,089	
9-letter words	110,753	227	607	235	864	975	

Figure 14: Statistical analysis of the South African text corpus in WordSmith Tools

Figure 15 below represents the statistical analysis of the Botswana text corpus in WordSmith consisting over 960,040 tokens.



N	1	2	3	4	5	6
Text File	OVERALL	UTLWAT~1.TXT	TLHOWA~1.TXT	SETSWA~4.TXT	SETSWA~3.TXT	SETSWA~2.TXT
Bytes	5,785,661	47,091	37,395	45,330	44,587	78,755
Tokens	960,040	8,291	6,522	7,043	6,752	11,307
Types	49,182	1,429	1,330	1,369	1,859	2,839
Type/Token Ratio	5.12	17.24	20.39	19.44	27.53	25.11
Standardised Type/Token	35.80	31.84	33.45	35.10	43.08	43.75
Ave. Word Length	3.76	3.62	3.59	3.97	3.94	3.95
Sentences	42,386	417	308	71	126	221
Sent length	20.01	18.99	20.26	57.73	34.09	38.72
sd. Sent. Length	23.48	13.42	13.01	54.93	30.98	48.45
Paragraphs	8,167	60	63	84	75	135
Para. length	115.59	132.63	101.97	83.38	86.23	83.70
sd. Para. length	751.96	220.06	185.87	163.83	154.29	159.21
Headings	0	0	0	0	0	0
Heading length						
sd. Heading length						
1-letter words	162,663	1,283	1,138	1,215	1,249	1,997
2-letter words	308,993	3,014	2,316	1,982	1,879	3,292
3-letter words	67,974	524	417	582	539	878
4-letter words	103,088	1,004	697	797	622	1,042
5-letter words	73,658	650	461	613	570	927
6-letter words	88,794	644	519	582	646	1,106
7-letter words	47,805	359	360	311	362	616
8-letter words	47,463	381	240	355	379	612
9-letter words	21,961	173	135	257	196	312

Figure 15: Statistical analysis of the Botswana text corpus in the WordSmith Tools

When comparing the statistical analysis of the South African text corpus and the Botswana text corpus as shown in Figures 14 and 15, the South African text corpus seems to be 4,69 larger than the Botswana text corpus.

Tables 12 and 13 below show the comparison between the South African written corpus and the Botswana text corpora in word lists consisting of the top 100 most frequently used words with the rankings on the left. The study is done to determine the relationship between the ranks and the frequency of a word.

Table 12: Comparison of the top 100 items between the South African written corpus and the Botswana text corpus

South African written corpus **v/s** **Botswana corpus**

Rank	Word	Rank	Word
1	**A**	1	A
2	**GO**	2	GO
3	**LE**	3	LE
4	**E**	4	E
5	**O**	5	O
6	**BA**	6	BA
7	**KA**	7	KA
8	**KE**	8	KE
10	**MO**	9	MO
9	**YA**	10	YA
11	**GA**	11	GA
15	**NE**	12	NE
12	**FA**	13	FA
13	**RE**	14	RE
14	**SE**	15	SE
16	**DI**	16	DI
18	**GORE**	17	GORE
19	**SA**	18	SA
17	**WA**	19	WA
21	**KWA**	20	KWA
25	**BO**	21	BO
20	**TSA**	22	TSA
24	**TSE**	23	TSE
23	**MME**	24	MME
22	**TLA**	25	TLA
55	**I**	26	I
29	**NNA**	27	NNA
26	**LA**	28	LA
32	**FELA**	29	FELA
28	**YO**	30	YO
30	**BONA**	31	BONA
27	**GAGWE**	32	GAGWE
33	**NA**	33	NA
35	**ITSE**	34	ITSE
34	**NTSE**	35	NTSE
41	**BONE**	36	BONE
39	**MOTHO**	37	MOTHO
49	**ENE**	38	ENE
42	**BATHO**	39	BATHO
75	**BANA**	42	BANA
36	**JAAKA**	43	JAAKA
31	**LO**	44	LO
48	**TSWA**	45	TSWA
68	**NGWANA**	46	NGWANA
58	**ENG**	47	ENG
45	**TENG**	48	TENG
43	**MONGWE**	49	MONGWE

38	**NENG**	50	NENG
37	**JWA**	51	JWA
57	**MONNA**	52	MONNA
46	**BUA**	53	BUA
89	**KANA**	55	KANA
64	**RILE**	56	RILE
52	**JALO**	58	JALO
56	**THATA**	60	THATA
87	**JANG**	62	JANG
71	**NAKO**	63	NAKO
60	**SETSE**	64	SETSE
50	**DIRA**	65	DIRA
53	**MORAGO**	66	MORAGO
69	**BATLA**	67	BATLA
76	**MOSADI**	69	MOSADI
70	**TSENA**	70	TSENA
74	**LETSATSI**	72	LETSATSI
47	**PELE**	73	PELE
78	**TSAYA**	74	TSAYA
44	**GAGO**	77	GAGO
40	**ME**	78	ME
59	**UTLWA**	79	UTLWA
92	**RAYA**	80	RAYA
79	**TSAMAYA**	81	TSAMAYA
80	**TLE**	82	TLE
83	**SENGWE**	83	SENGWE
85	**SENTLE**	84	SENTLE
73	**KGOTSA**	85	KGOTSA
62	**WENA**	87	WENA
67	**PELO**	88	PELO
72	**GAPE**	89	GAPE
94	**KAE**	91	KAE
63	**RONA**	92	RONA
77	**JAANONG**	94	JAANONG
84	**BILE**	96	BILE
86	DILO	***101***	***DILO***
90	RATA	***102***	***RATA***
91	MATLHO	***104***	***MATLHO***
65	NTLHA	***105***	***NTLHA***
97	NNGWE	***106***	***NNGWE***
96	JAANA	***107***	***JAANA***
81	MAFOKO	***117***	***MAFOKO***
88	GODIMO	***123***	***GODIMO***
98	JO	***135***	***JO***
100	TSOTLHE	***153***	***TSOTLHE***
51	KGOSI	***171***	***KGOSI***
66	YONA	***212***	***YONA***
82	ENA	***228***	***ENA***
99	IWA	***260***	***IWA***
54	GONNE	***275***	***GONNE***
95	MODIMO	***278***	***MODIMO***
93	BE	***405***	***BE***

If one now compares the South African written corpus and the Botswana text corpus in Table 12, focusing on the ranking of the top 100 items in both corpora. An important observation stemming from the comparison is that both the two corpora appear to have typical features of the same ranking orders with 13% of the items common in both the South African and the Botswana text corpora. However, it is important to note that 83% of the items in the Botswana text corpus are retained while 17% of the items fall outside the top 100 items of the South African written corpus. It is important to state that all the 17 ousted items fall very close in the range 101-405. Conclusion can thus far be drawn that both the South African and the Botswana text corpora are much closer to each other and that there are still considerable ranking overlaps.

Table 13: Comparison of the top 100 items between the Botswana text corpus and the South African written corpus

Botswana corpus **v/s** **South African written corpus**

Rank	Word	Rank	Word
1	**A**	1	A
2	**GO**	2	GO
3	**LE**	3	LE
4	**E**	4	E
5	**O**	5	O
6	**BA**	6	BA
7	**KA**	7	KA
8	**KE**	8	KE
10	**YA**	9	YA
9	**MO**	10	MO
11	**GA**	11	GA
13	**FA**	12	FA
14	**RE**	13	RE
15	**SE**	14	SE
12	**NE**	15	NE
16	**DI**	16	DI
19	**WA**	17	WA
17	**GORE**	18	GORE
18	**SA**	19	SA
22	**TSA**	20	TSA
20	**KWA**	21	KWA
25	**TLA**	22	TLA
24	**MME**	23	MME
23	**TSE**	24	TSE
21	**BO**	25	BO
28	**LA**	26	LA
32	**GAGWE**	27	GAGWE

30	**YO**	28	YO
27	**NNA**	29	NNA
31	**BONA**	30	BONA
44	**LO**	31	LO
29	**FELA**	32	FELA
33	**NA**	33	NA
35	**NTSE**	34	NTSE
34	**ITSE**	35	ITSE
43	**JAAKA**	36	JAAKA
51	**JWA**	37	JWA
50	**NENG**	38	NENG
37	**MOTHO**	39	MOTHO
78	**ME**	40	ME
36	**BONE**	41	BONE
39	**BATHO**	42	BATHO
49	**MONGWE**	43	MONGWE
77	**GAGO**	44	GAGO
48	**TENG**	45	TENG
53	**BUA**	46	BUA
73	**PELE**	47	PELE
45	**TSWA**	48	TSWA
38	**ENE**	49	ENE
65	**DIRA**	50	DIRA
58	**JALO**	52	JALO
66	**MORAGO**	53	MORAGO
26	**I**	55	I
60	**THATA**	56	THATA
52	**MONNA**	57	MONNA
47	**ENG**	58	ENG
79	**UTLWA**	59	UTLWA
64	**SETSE**	60	SETSE
87	**WENA**	62	WENA
92	**RONA**	63	RONA
56	**RILE**	64	RILE
88	**PELO**	67	PELO
46	**NGWANA**	68	NGWANA
67	**BATLA**	69	BATLA
70	**TSENA**	70	TSENA
63	**NAKO**	71	NAKO
89	**GAPE**	72	GAPE
85	**KGOTSA**	73	KGOTSA
72	**LETSATSI**	74	LETSATSI
42	**BANA**	75	BANA
69	**MOSADI**	76	MOSADI
94	**JAANONG**	77	JAANONG
74	**TSAYA**	78	TSAYA
81	**TSAMAYA**	79	TSAMAYA
82	**TLE**	80	TLE
83	**SENGWE**	83	SENGWE
96	**BILE**	84	BILE
84	**SENTLE**	85	SENTLE
62	**JANG**	87	JANG
55	**KANA**	89	KANA
80	**RAYA**	92	RAYA
91	**KAE**	94	KAE

95	TIRO	*104*	*TIRO*
76	JA	*109*	*JA*
100	MADI	*112*	*MADI*
93	TOTA	*116*	*TOTA*
40	L	*119*	*L*
86	GONE	*125*	*GONE*
90	B	*134*	*B*
97	YONE	*140*	*YONE*
61	TWE	*142*	*TWE*
68	C	*150*	*C*
57	T	*156*	*T*
75	R	*194*	*R*
59	MMA	*200*	*MMA*
99	THUSA	*239*	*THUSA*
71	BAITHUTI	*410*	*BAITHUTI*
54	TLAA	*614*	*TLAA*

If one follows the same approach of comparing the top 100 items in the Botswana text corpus and the South African written corpus, one concludes that 84% (one item was not considered) of the items in the South African written corpus are retained. An analysis indicates that 12% of the retained items share the same ranking orders. It is also important to note that only 16% of the items in the South African written corpus fall outside the top 100 items in the Botswana text corpus but are still regarded as the most frequent used items in the South African written corpus. Now if one considers the ousted items in Table 14, one concludes that 16 ousted items fall very close in the range 101-410 with the exception of *tlaa* falling slightly far the top 100 items. Conclusion can thus far be drawn that there is very little difference between the two corpora and thus they are very closely related.

When the two word lists are compared in terms of their frequency items, the South African written corpus seems to be 4,36 times larger than the Botswana text corpus. According to Scott and Tribble (1996:23–24), the word list consists of different high, medium and low frequency items. The function words (high) are brought to the top. It is clear from both word lists that function words occur frequently and that frequencies are Zipfian i.e. very rapid descend with approximately half of the types hapax legomena as indicated in Figure 16 below.

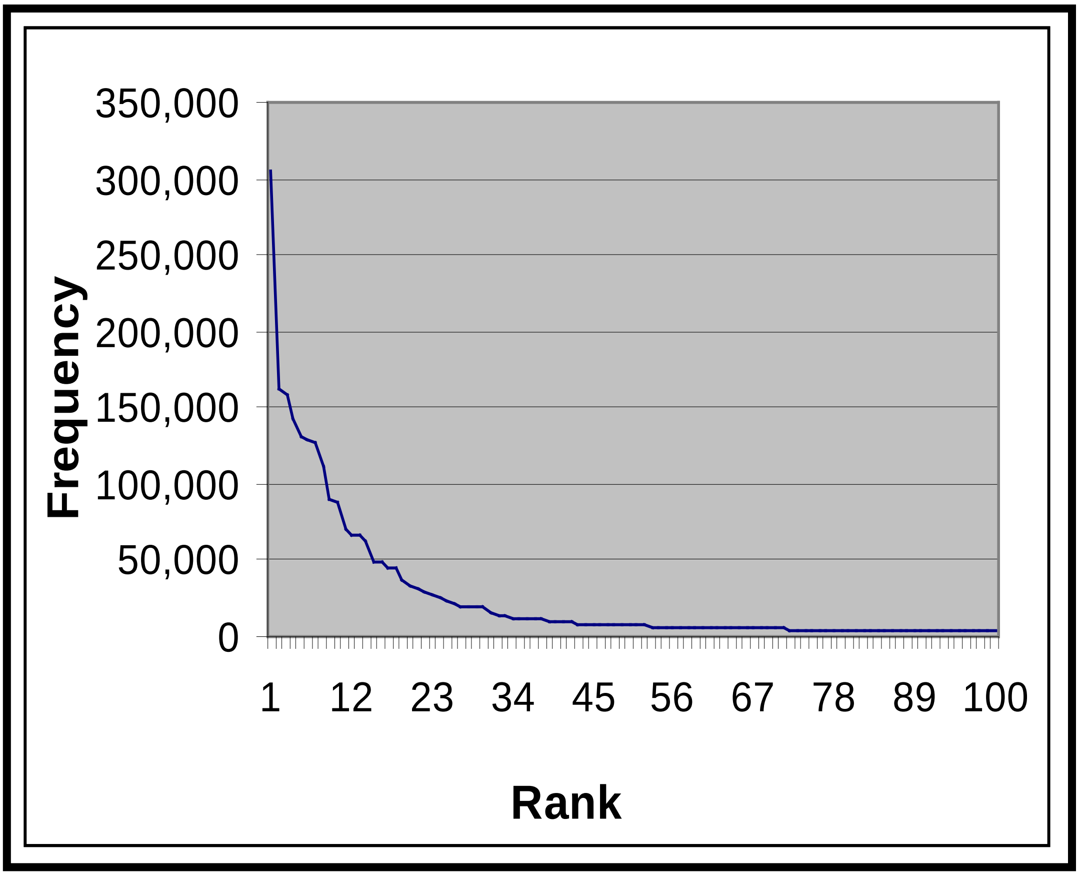
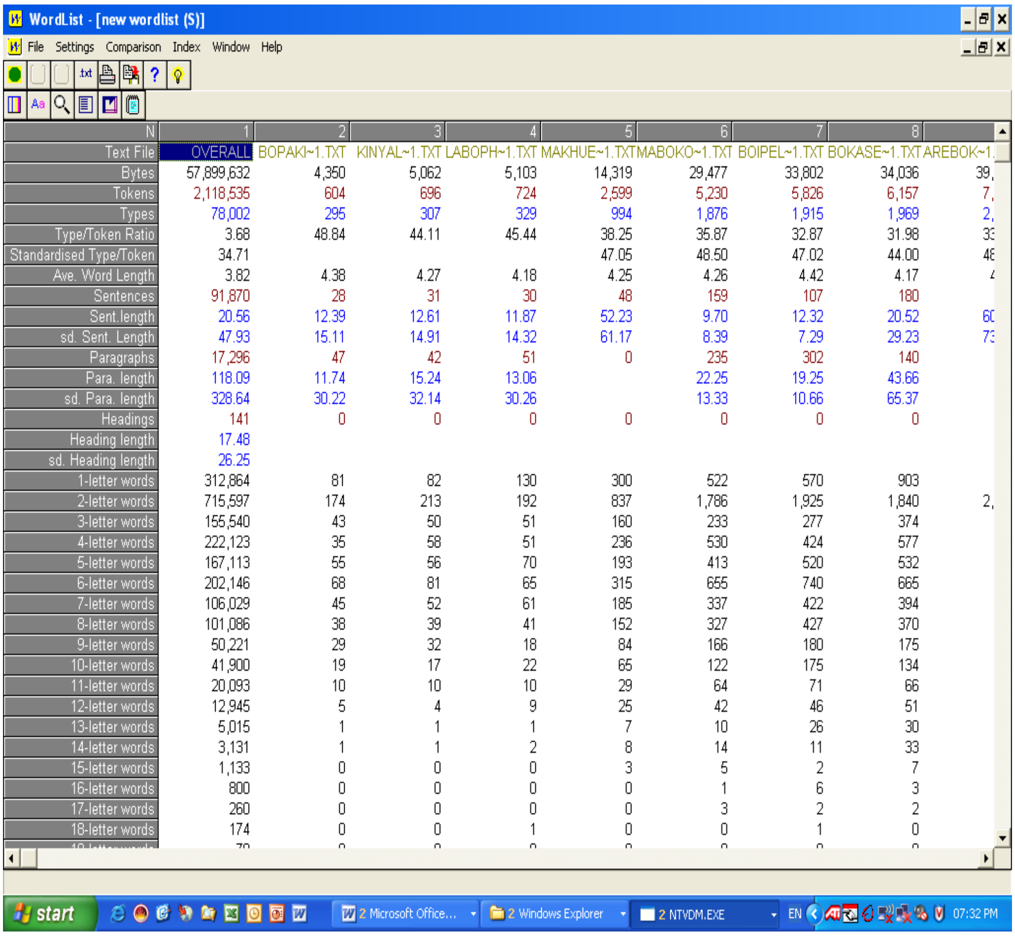


Figure 16: A rapid frequency decline in the top 100 words

3.11 Testing the stability of the Setswana corpus

The aim here is to assess the stability of growing organic corpora for Setswana. De Schryver and Prinsloo (2000:3) describe the aim of testing the stability of growing organic corpora for the Bantu languages as a process that entails a series of stability tests in order to determine whether or not substantial enlargement of corpora or even doubling or tripling of their size, will substantiate conclusions which were drawn during the earlier stages in the development of the corpora. In the following section this model will be applied to Setswana. The corpus was divided into two different sections and the sections were compared with each other.

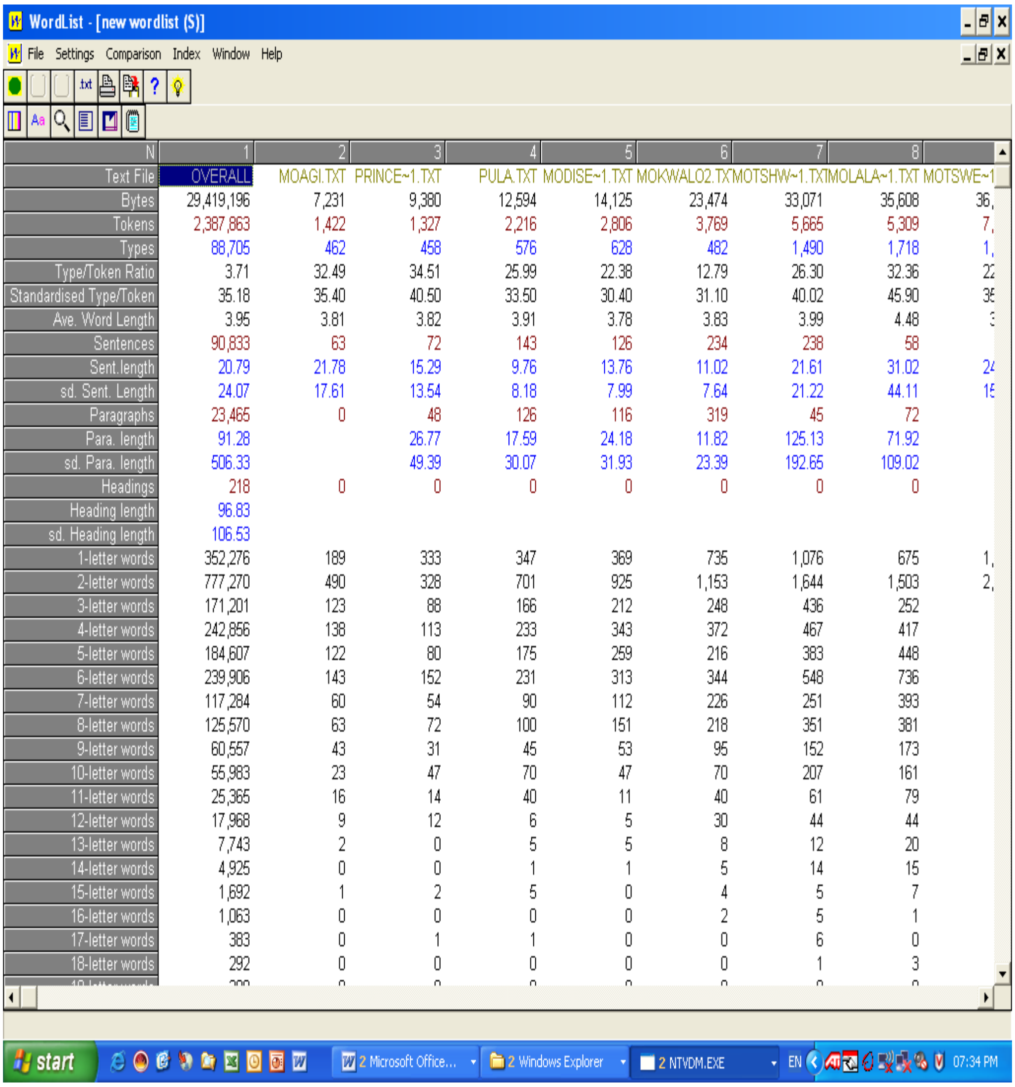
The statistical analysis drawn from the Pretoria Setswana text Corpus (PSetC) will now be used to monitor the growing stages of the Setswana corpus. The first phase (PSetC-Phase1) was doubled to contain 2118, 535 tokens (PSetC-Half1). Consider figure 17 below:



N	1	2	3	4	5	6	7	8	
Text File	OVERALL	BOPAKI~1.TXT	KINYAL~1.TXT	LABOPH~1.TXT	MAKHUE~1.TXT	MABOKO~1.TXT	BOIPEL~1.TXT	BOKASE~1.TXT	AREBOK~1.
Bytes	57,899,632	4,350	5,062	5,103	14,319	29,477	33,802	34,036	39,
Tokens	2,118,535	604	696	724	2,599	5,230	5,826	6,157	7,
Types	78,002	295	307	329	994	1,876	1,915	1,969	2,
Type/Token Ratio	3.68	48.84	44.11	45.44	38.25	35.87	32.87	31.98	33
Standardised Type/Token	34.71				47.05	48.50	47.02	44.00	46
Ave. Word Length	3.82	4.38	4.27	4.18	4.25	4.26	4.42	4.17	4
Sentences	91,870	28	31	30	48	159	107	180	
Sent.length	20.56	12.39	12.61	11.87	52.23	9.70	12.32	20.52	60
sd. Sent. Length	47.93	15.11	14.91	14.32	61.17	8.39	7.29	29.23	73
Paragraphs	17,296	47	42	51	0	235	302	140	
Para. length	118.09	11.74	15.24	13.06		22.25	19.25	43.66	
sd. Para. length	328.64	30.22	32.14	30.26		13.33	10.66	65.37	
Headings	141	0	0	0	0	0	0	0	
Heading length	17.48								
sd. Heading length	26.25								
1-letter words	312,864	81	82	130	300	522	570	903	
2-letter words	715,597	174	213	192	837	1,786	1,925	1,840	2,
3-letter words	155,540	43	50	51	160	233	277	374	
4-letter words	222,123	35	58	51	236	530	424	577	
5-letter words	167,113	55	56	70	193	413	520	532	
6-letter words	202,146	68	81	65	315	655	740	665	
7-letter words	106,029	45	52	61	185	337	422	394	
8-letter words	101,086	38	39	41	152	327	427	370	
9-letter words	50,221	29	32	18	84	166	180	175	
10-letter words	41,900	19	17	22	65	122	175	134	
11-letter words	20,093	10	10	10	29	64	71	66	
12-letter words	12,945	5	4	9	25	42	46	51	
13-letter words	5,015	1	1	1	7	10	26	30	
14-letter words	3,131	1	1	2	8	14	11	33	
15-letter words	1,133	0	0	0	3	5	2	7	
16-letter words	800	0	0	0	0	1	6	3	
17-letter words	260	0	0	0	0	3	2	2	
18-letter words	174	0	0	1	0	0	1	0	

Figure 17: Statistical analysis of the Pretoria Setswana Corpus – Phase 1 (PSetC-Half1)

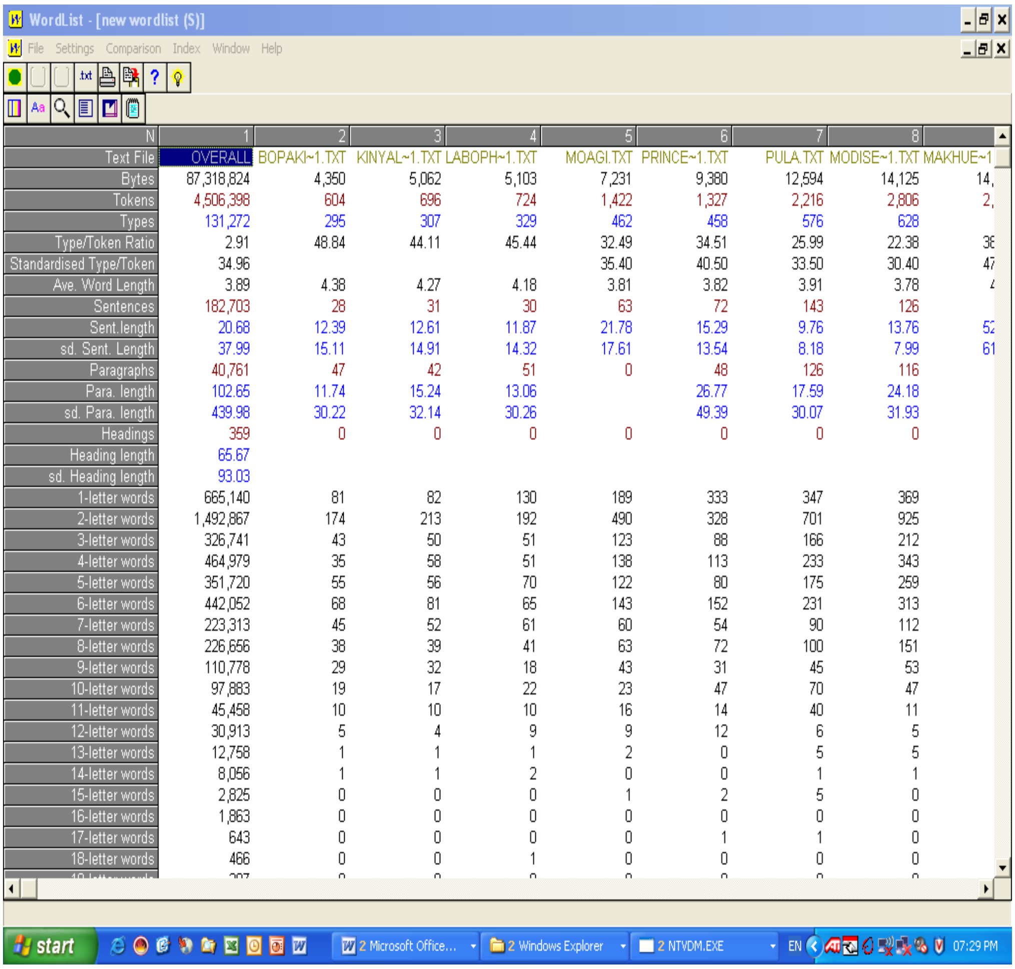
A different corpus was subsequently compiled from a half corpus of 2,387,863 (PSetC-Half 2) tokens. Consider the statistical analysis in figure 18 below:


WordList - [new wordlist (S)]
File Settings Comparison Index Window Help
start
2 Microsoft Office...
2 Windows Explorer
2 NTVDM.EXE
EN
07:34 PM

N	1	2	3	4	5	6	7	8	
Text File	OVERALL	MOAGI.TXT	PRINCE~1.TXT	PULA.TXT	MODISE~1.TXT	MOKWALO2.TX	MOTSHW~1.TX	MOLALA~1.TXT	MOTSWE~1
Bytes	29,419,196	7,231	9,380	12,594	14,125	23,474	33,071	35,608	36,
Tokens	2,387,863	1,422	1,327	2,216	2,806	3,769	5,665	5,309	7,
Types	88,705	462	458	576	628	482	1,490	1,718	1,
Type/Token Ratio	3.71	32.49	34.51	25.99	22.38	12.79	26.30	32.36	22
Standardised Type/Token	35.18	35.40	40.50	33.50	30.40	31.10	40.02	45.90	35
Ave. Word Length	3.95	3.81	3.82	3.91	3.78	3.83	3.99	4.48	3
Sentences	90,833	63	72	143	126	234	238	58	
Sent.length	20.79	21.78	15.29	9.76	13.76	11.02	21.61	31.02	24
sd. Sent. Length	24.07	17.61	13.54	8.18	7.99	7.64	21.22	44.11	15
Paragraphs	23,465	0	48	126	116	319	45	72	
Para. length	91.28		26.77	17.59	24.18	11.82	125.13	71.92	
sd. Para. length	506.33		49.39	30.07	31.93	23.39	192.65	109.02	
Headings	218	0	0	0	0	0	0	0	
Heading length	96.83								
sd. Heading length	106.53								
1-letter words	352,276	189	333	347	369	735	1,076	675	1,
2-letter words	777,270	490	328	701	925	1,153	1,644	1,503	2,
3-letter words	171,201	123	88	166	212	248	436	252	
4-letter words	242,856	138	113	233	343	372	467	417	
5-letter words	184,607	122	80	175	259	216	383	448	
6-letter words	239,906	143	152	231	313	344	548	736	
7-letter words	117,284	60	54	90	112	226	251	393	
8-letter words	125,570	63	72	100	151	218	351	381	
9-letter words	60,557	43	31	45	53	95	152	173	
10-letter words	55,983	23	47	70	47	70	207	161	
11-letter words	25,365	16	14	40	11	40	61	79	
12-letter words	17,968	9	12	6	5	30	44	44	
13-letter words	7,743	2	0	5	5	8	12	20	
14-letter words	4,925	0	0	1	1	5	14	15	
15-letter words	1,692	1	2	5	0	4	5	7	
16-letter words	1,063	0	0	0	0	2	5	1	
17-letter words	383	0	1	1	0	0	6	0	
18-letter words	292	0	0	0	0	0	1	3	

Figure 18: Statistical analysis of the Pretoria Setswana Corpus - Phase 2 (PSetC-Half2*)*

The combination size of the PSetC Phase1 (PSetC-Half1) and Phase 2 (PSetC-Half2) gives the size of 4506,398 tokens (PSetC-Sum). Consider Figure 19 below:



N	1	2	3	4	5	6	7	8	
Text File	OVERALL	BOPAKI~1.TXT	KINYAL~1.TXT	LABOPH~1.TXT	MOAGI.TXT	PRINCE~1.TXT	PULA.TXT	MODISE~1.TXT	MAKHUE~1
Bytes	87,318,824	4,350	5,062	5,103	7,231	9,380	12,594	14,125	14,
Tokens	4,506,398	604	696	724	1,422	1,327	2,216	2,806	2,
Types	131,272	295	307	329	462	458	576	628	
Type/Token Ratio	2.91	48.84	44.11	45.44	32.49	34.51	25.99	22.38	38
Standardised Type/Token	34.96				35.40	40.50	33.50	30.40	47
Ave. Word Length	3.89	4.38	4.27	4.18	3.81	3.82	3.91	3.78	4
Sentences	182,703	28	31	30	63	72	143	126	
Sent.length	20.68	12.39	12.61	11.87	21.78	15.29	9.76	13.76	52
sd. Sent. Length	37.99	15.11	14.91	14.32	17.61	13.54	8.18	7.99	61
Paragraphs	40,761	47	42	51	0	48	126	116	
Para. length	102.65	11.74	15.24	13.06		26.77	17.59	24.18	
sd. Para. length	439.98	30.22	32.14	30.26		49.39	30.07	31.93	
Headings	359	0	0	0	0	0	0	0	
Heading length	65.67								
sd. Heading length	93.03								
1-letter words	665,140	81	82	130	189	333	347	369	
2-letter words	1,492,867	174	213	192	490	328	701	925	
3-letter words	326,741	43	50	51	123	88	166	212	
4-letter words	464,979	35	58	51	138	113	233	343	
5-letter words	351,720	55	56	70	122	80	175	259	
6-letter words	442,052	68	81	65	143	152	231	313	
7-letter words	223,313	45	52	61	60	54	90	112	
8-letter words	226,656	38	39	41	63	72	100	151	
9-letter words	110,778	29	32	18	43	31	45	53	
10-letter words	97,883	19	17	22	23	47	70	47	
11-letter words	45,458	10	10	10	16	14	40	11	
12-letter words	30,913	5	4	9	9	12	6	5	
13-letter words	12,758	1	1	1	2	0	5	5	
14-letter words	8,056	1	1	2	0	0	1	1	
15-letter words	2,825	0	0	0	1	2	5	0	
16-letter words	1,863	0	0	0	0	0	0	0	
17-letter words	643	0	0	0	0	1	1	0	
18-letter words	466	0	0	1	0	0	0	0	

Figure 19: Statistical analysis of the Pretoria Setswana Corpus - Phase 1 (PSetC-Half1) and phase2 (PSetC-Half2)

If one compares the items in the different rank ranges of PSetC-Half1, or respectively, of PSetC-Half2 to those in PSetC-Sum, one gets the results shown in figures 20 and 21 below.

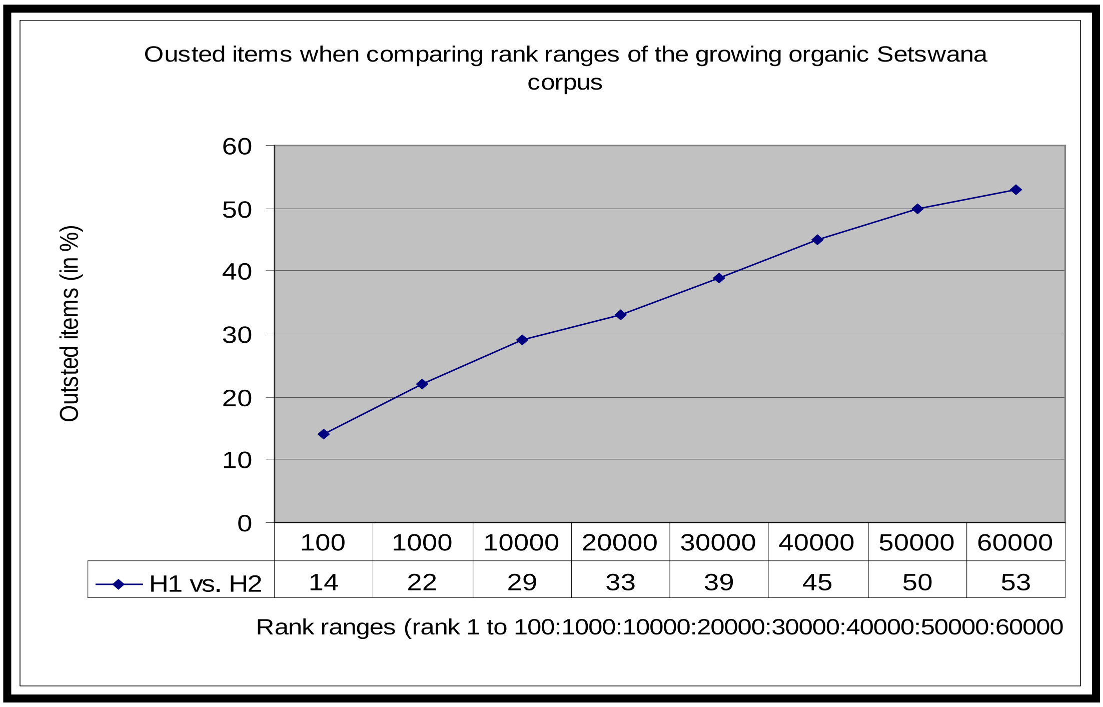


Figure 20: Items when comparing rank ranges of PSetC-Half1 with PSetC-Half2

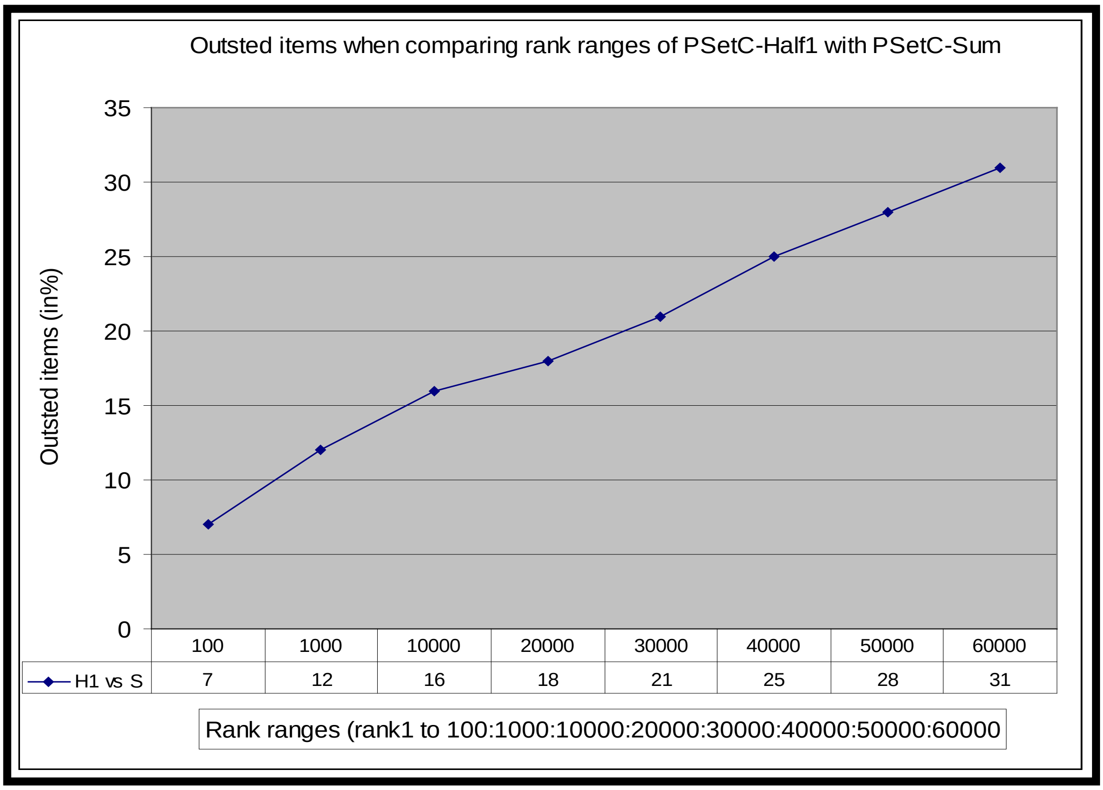
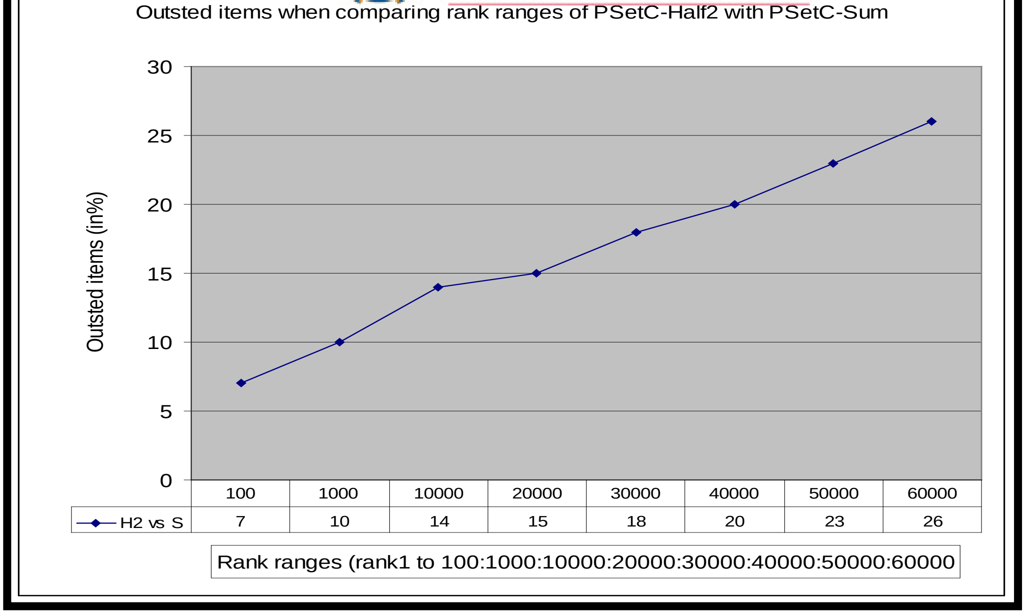


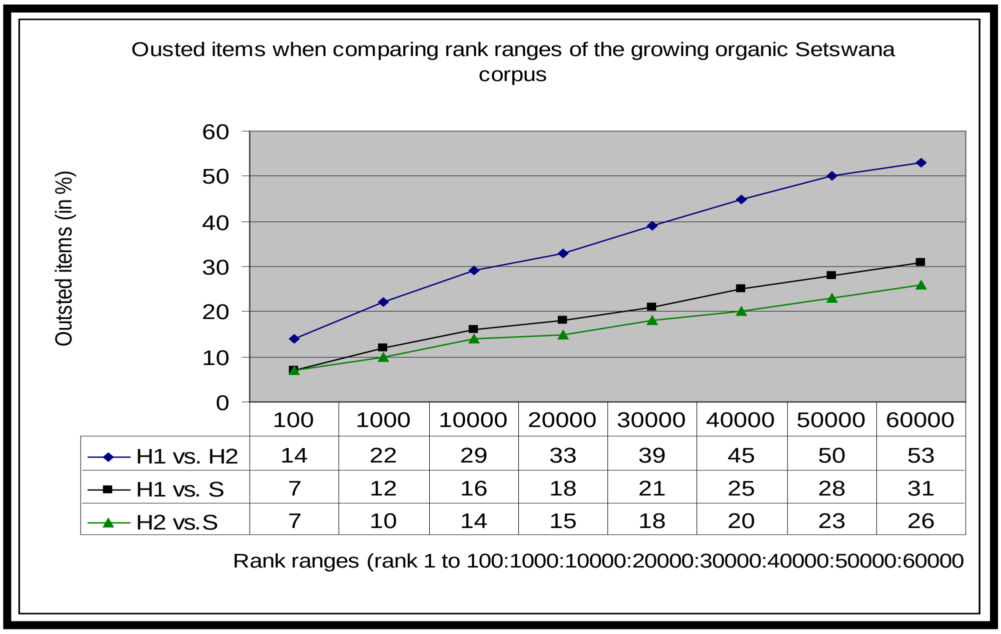
Figure 21: Items when comparing rank ranges of PSetC-Half1 with PSetC-Sum



	100	1000	10000	20000	30000	40000	50000	60000
H2 vs S	7	10	14	15	18	20	23	26

Figure 22: Ousted items when comparing rank ranges of PSetC-Half2 with PSetC-Sum

From the combined Setswana data presented in Figures 20, 21 and 22 above, stability conclusions for the growing Setswana corpus can be drawn in Figure 23



	100	1000	10000	20000	30000	40000	50000	60000
H1 vs. H2	14	22	29	33	39	45	50	53
H1 vs. S	7	12	16	18	21	25	28	31
H2 vs.S	7	10	14	15	18	20	23	26

Figure 23: Ousted items when comparing rank ranges of the growing organic Setswana corpus

The top graph represents the two unrelated Setswana corpora (H1 vs. H2). The second graph represents the Setswana-Half1 in comparison to the Setswana-Sum (H1 vs. S).

The bottom graph represents items in different rank ranges of the Setswana-Half2 in comparison to the Setswana-Sum (Half2 vs. S). Conclusion can thus far be made that the two bottom graphs reveal the stability of the growing Setswana corpus. Consider Table 14 below as a summary of Figure 23.

Table 14: Summary of SetH1 versus SetH2

SetH1 versus SetH2	In common	ousted	%
100	86	14	14
1000	777	223	22
10000	7082	2918	29
20000	13359	6641	33
30000	18323	11677	39
40000	22045	17955	45
50000	25016	24984	50
60000	27997	32003	53
SetH1 versus SetSum			
100	93	7	7
1000	879	121	12
10000	8432	1568	16
20000	16374	3626	18
30000	23683	6317	21
40000	30128	9872	25
50000	35813	14187	28
60000	41413	18587	31
SetH2 versus SetSum			
100	93	7	7
1000	898	102	10
10000	8648	1352	14
20000	16966	3034	15
30000	24640	5360	18
40000	31894	8106	20
50000	38283	11717	23
60000	44642	15358	26

3.12 Conclusion

In this chapter we demonstrated the value of the corpora such as the Brown corpus and the Lancaster-Oslo/ Bergen (LOB) corpus as a corpus-based approach to the compilation of a corpus-based Setswana dictionary. We have shown how COBUILD was used to address a number of issues geared towards achieving representativeness and balance corpora in the corpus design, as well as aspects relating to the size of a corpus. This chapter was mostly concerned with the collection of the oral corpus. It provides detailed account of how the oral corpus can be of value to the dictionary compilers. We have also seen how projects were used to compile the oral data. The process of oral data collection has been effectively highlighted covering issues of keyness function. Important concepts in keyness function such as positive keyness and negative keyness were also contrasted and graphically represented. Although in the oral corpora we did not encounter serious language deficiences, we highlighted serious typical differences of language use between oral and written corpora. They were selected from respondents with different backgrounds.

Illustrated contrasts between both the South African spoken corpus and the South African oral corpus and the South African text corpora and the Botswana text corpora have been demonstrated using the WordSmith Tools. We have seen how studies are being conducted to monitor the stability of the growing organic corpora for Setswana using sophisticated computer interface packages, called ‘corpus query tools’ to analyse a corpus in various ways. We observed stabilities of growing organic Setswana corpora that support the method of compilation which has been followed so far in producing dictionaries for Sepedi.

In conclusion, Prinsloo and De Schryver (2001:39) suggest that corpus compilers need to constantly monitor their growing corpora in accordance with the methods and criteria outlined in their article. This is aimed necessary as a means avoiding situation, where a substantial, yet blind enlargement of the corpus build result in severe skewing.

Chapter 4

The macrostructure

4.1 Introduction

The aim of this chapter is firstly to critically evaluate and analyse the macrostructure of existing Setswana dictionaries with a focus on the compilation of and the deficiencies in the lemma lists. Typical macrostructural inconsistencies existing in Setswana dictionaries will be highlighted. The extent of the inconsistencies will generally show how the respective Setswana dictionaries succeed or fail to treat the most commonly used words. The focus will mainly be on the following macrostructural aspects: inconsistencies regarding the lemmatization of nouns and verbal derivations, imbalances regarding the alphabetical stretches, lemmatization approaches, lemmatization strategies, lemmatization traditions and grammatical aspects regarding the lemmatization problems of the noun prefixes of classes 5 and 11; the absence of noun inflections and the inconsistencies regarding the lemmatization of homonyms and the absence of tonal indication.

Secondly, the chapter will demonstrate how corpus query tools can be used to generate alphabetical word lists and frequency lists reflecting the overall counts or specific words or words in context. Thereafter follows the plotting of data to indicate the relationship between rank and the frequency of tokens. Thirdly, the chapter will demonstrate how the Setswana dictionaries handle the treatment of dialectical words. Finally, each section dealing with the inconsistencies will conclude with suggestions for the improvement of the respective Setswana dictionaries by means of a corpus-based macrostructure.

According to Prinsloo and Gouws (1996:103), the lexicographer for African languages must find lemmatization strategies that result in a user-friendly end product. It is important for the lexicographer to find a sound balance in terms of the selection of lemmata for words likely to be looked up by the target users. Martin et al. (1983:81-82, 87) state that:

> "The decision what to include in the dictionary still has to be made by the lexicographer himself, however, and this depends in turn upon the nature and size of the dictionary and its intended users. In this respect lemmatized frequency-lists can be a further help… We have reached a stage where co-operation between man and machine is useful and perhaps indispensable in making better dictionaries"

When examining the macrostructure of the existing Setswana dictionaries in comparison to word lists culled from a Setswana corpus, one can easily determine the typical macrostructural inconsistencies that need to be rectified through the electronic corpora. Gouws (1990:55) states:

> "Lexicographical activities on the various indigenous African Languages […have] resulted in a wide range of dictionaries. Unfortunately, the majority of the dictionaries are the products of limited efforts not reflecting a high standard of lexicographical achievement."

4.2 Typical macrostructural inconsistencies existing in Setswana dictionaries

There is no dictionary that can be comprehensive enough to give a balanced account of a representative selection of a lexicon. However selection of lexical items to be included as lemmas cannot be done in an arbitrary way, but have to comply with the lexicographical standards rooted in a sound theory (cf. Prinsloo and Gouws 1995:1). The analysis of dictionaries in randomly selected alphabetical stretches or sections of alphabetical stretches in Table 15 reveals the importance of the utilization of corpora during the creation of a dictionary's lemma-sign list.

Table 15: Comparison of the macrostructure between the fixed points *rabbit* and *rally* in various dictionaries

S.A. Oxford School dictionary. (Oxford university Press 2004:352-357)	Major Dictionary. (Eksteen, 1997:1238-1241)	Setswana-English Dictionary (Brown, 1964:512-513)	Dikišinari ya Setswana. (Snyman et. al, 1990:313)	Setswana-English-Setswana dictionary. (Matumo, 1993:590)	English- Sepedi Kriel 1976
rabbit	rabbit	rabbit	rabbit	rabbit	rabbit
race	race	race	race	race	race
racialism	____	____	____	____	____
racism	____	____	____	____	racism
racist	racist	____	____	____	____
rack	rack	____	____	rack	rack
radial	radial	____	____	____	____
radiate	radiate	radiate	radiate	radiate	radiate
radiator	radiator	____	____	____	____
radical	radical	____	____	____	____
radio	radio	____	____	____	radio
radiographer	radiographer	____	____	____	____
radiology	____	____	____	____	radiology
radish	radish	radish	____	____	radish
radius	radius	____		____	radius
raffle	raffle	____	radius	____	____
rag	rag	rag	rag	rag	rag
raid	raid	raid	____	raid	raid
rail	rail	rail	rail	rail	rail
rain	rain	rain	rain	rain	rain
raise	raise	raise	raise	raise	raise
rake	rake	rake	____	rake	rake
rally	rally	____	____	rally	rally

In Table 15, six dictionaries are compared and viewed over 23 items in the alphabetical stretch *rabbit* to *rally*. The two English dictionaries Oxford University Press (2004) and Eksteen (1997) are used as a base to indicate the macrostructural inconsistencies existing in Setswana dictionaries. Brown (1964) treats 10 items, Snyman et al (1990) treat 8 items, Matumo (1993) treats 11 and Kriel (1976) treats 16 items. It is important to note that words most likely to be consulted by the target users have been left out in most of the Setswana dictionaries as reflected in Table 15. The number of the lexical items and the absence of the treatment of most commonly used words like *racial, racist, radial, radiator, radical, radiology, radius* etc. in Setswana dictionaries, prove the urgent need of corpus utilization of word frequency counts to compile a lemmatized frequency list.

For the revision of existing Setswana dictionaries, frequency lists can play a vital role in ascertaining that frequently used words are not accidentally omitted, and, on the other hand, that dictionary space is not occupied by articles of lemmas unlikely to be looked for by the target users.

Consider the second example where inconsistency regarding the entering of derivations is evident in Setswana monolingual dictionaries.

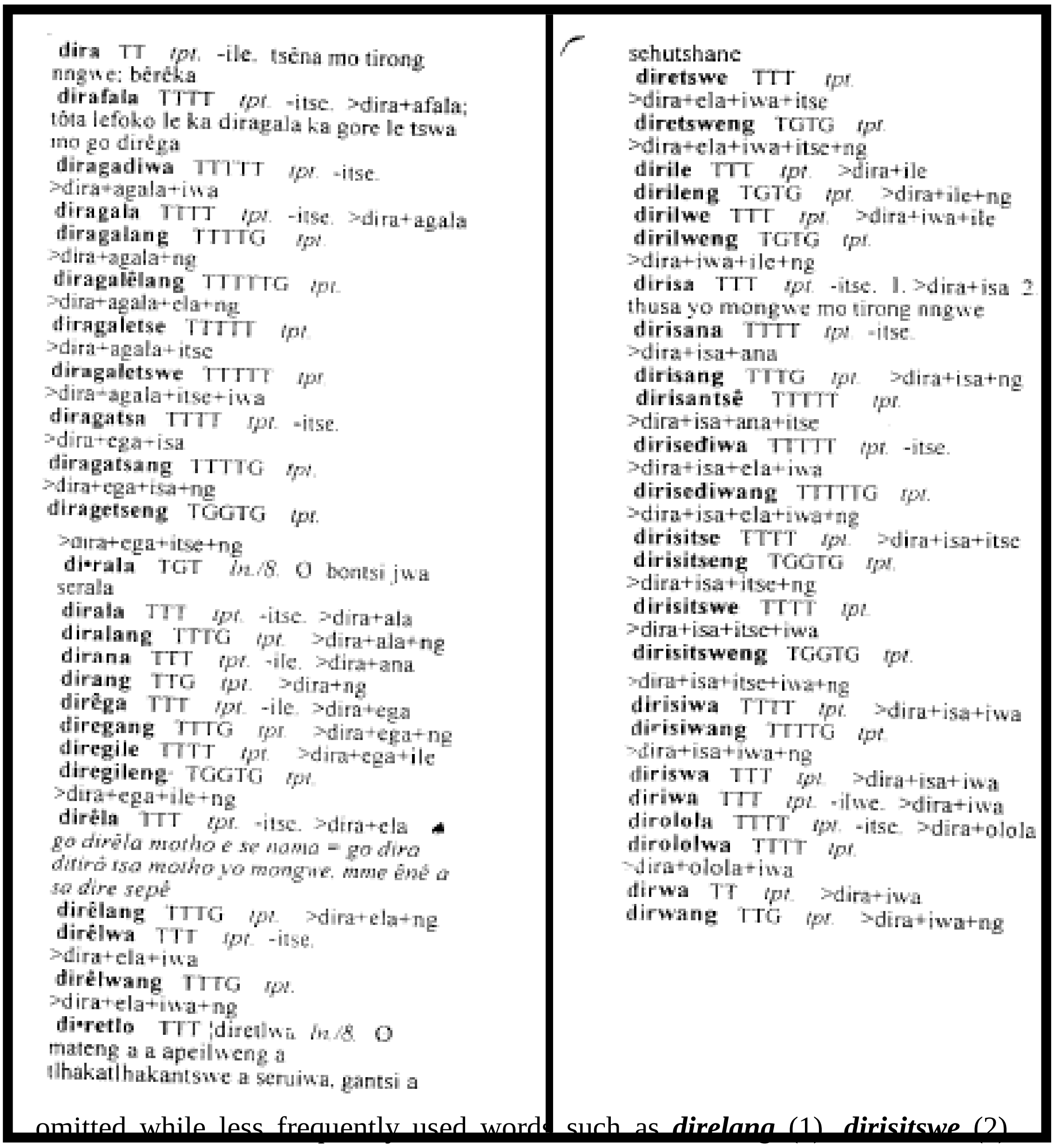
Example 3: *Reka (buy)* (672)

rekegela **(**36), *rekela* (125), ***rekelana* (0), *rekelane* (0), *rekelwa* (8),** *reketswe* (15), *rekile* (150), *rekileng* (26), *rekilwe* (34), *rekisa* (280), *rekisang* (32), ***rekiseditswe* (2)**, *rekisetsa* (40), *rekisitswe* (14), *rekisiwa* (43), ***rekiwa* (5),** *rekolola* (35), *rekwa* (69), <u>*reketse*</u> (35), <u>*rekang* </u>(37), <u>*reke*</u> (79), <u>*rekegela*</u> (36), <u>*rekegele*</u> (35), <u>*rekele*</u> (20), <u>*rekise*</u> (30), <u>*rekileng*</u> (26), <u>*rekisang*</u> (32), <u>*rekisitse* </u>(21), <u>*rekwe*</u> (19).

From this example, the inconsistency regarding the entering of derivations in THAN is evident. It is difficult to justify the inclusion and the exclusion of the derivations of the verb *reka* (buy). Frequencies are given between brackets in example 3. Highly used underlined derivations such as <u>*reke*</u> (79), <u>*reketse*</u> (35), rekang (37), rekise (30), rekisang (32) etc. have been omitted while less frequently used derivations (bolded) such as ***rekelana*** (0), ***rekelane*** (0), ***rekelwa*** (8), ***rekiseditswe*** (2), ***rekiwa*** (5) etc. are entered.

Consider the third example of inconsistency in the THAN where highly frequently used derivatives of the verb *dira* are omitted or not lemmatised while less frequently used ones are lemmatized:

Example 4: THAN



dira TT *tpt.* -ile. tsêna mo tirong nngwe: bêrêka
dirafala TTTT *tpt.* -itse. >dira+afala; tôta lefoko le ka diragala ka gore le tswa mo go dirêga
diragadiwa TTTTT *tpt.* -itse. >dira+agala+iwa
diragala TTTT *tpt.* -itse. >dira+agala
diragalang TTTTG *tpt.* >dira+agala+ng
diragalêlang TTTTTG *tpt.* >dira+agala+ela+ng
diragaletse TTTTT *tpt.* >dira+agala+itse
diragaletswe TTTTT *tpt.* >dira+agala+itse+iwa
diragatsa TTTT *tpt.* -itse. >dira+ega+isa
diragatsang TTTTG *tpt.* >dira+ega+isa+ng
diragetseng TGGTG *tpt.* >dira+ega+itse+ng
di•rala TGT *In./8.* O bontsi jwa serala
dirala TTT *tpt.* -itse. >dira+ala
diralang TTTG *tpt.* >dira+ala+ng
dirana TTT *tpt.* -ile. >dira+ana
dirang TTG *tpt.* >dira+ng
dirêga TTT *tpt.* -ile. >dira+ega
diregang TTTG *tpt.* >dira+ega+ng
diregile TTTT *tpt.* >dira+ega+ile
diregileng TGGTG *tpt.* >dira+ega+ile+ng
dirêla TTT *tpt.* -itse. >dira+ela *go dirêla motho e se nama = go dira ditirô tsa motho yo mongwe. mme ênê a sa dire sepê*
dirêlang TTTG *tpt.* >dira+ela+ng
dirêlwa TTT *tpt.* -itse. >dira+ela+iwa
dirêlwang TTTG *tpt.* >dira+ela+iwa+ng
di•retlo TTT |diretlwa *In./8.* O mateng a a apeilweng a tlhakatlhakantswe a seruiwa, gantsi a sehutshane
diretswe TTT *tpt.* >dira+ela+iwa+itse
diretsweng TGTG *tpt.* >dira+ela+iwa+itse+ng
dirile TTT *tpt.* >dira+ile
dirileng TGTG *tpt.* >dira+ile+ng
dirilwe TTT *tpt.* >dira+iwa+ile
dirilweng TGTG *tpt.* >dira+iwa+ile+ng
dirisa TTT *tpt.* -itse. 1. >dira+isa 2. thusa yo mongwe mo tirong nngwe
dirisana TTTT *tpt.* -itse. >dira+isa+ana
dirisang TTTG *tpt.* >dira+isa+ng
dirisantsê TTTTT *tpt.* >dira+isa+ana+itse
dirisediwa TTTTT *tpt.* -itse. >dira+isa+ela+iwa
dirisediwang TTTTTG *tpt.* >dira+isa+ela+iwa+ng
dirisitse TTTT *tpt.* >dira+isa+itse
dirisitseng TGGTG *tpt.* >dira+isa+itse+ng
dirisitswe TTTT *tpt.* >dira+isa+itse+iwa
dirisitsweng TGGTG *tpt.* >dira+isa+itse+iwa+ng
dirisiwa TTTT *tpt.* >dira+isa+iwa
dirisiwang TTTTG *tpt.* >dira+isa+iwa+ng
diriswa TTT *tpt.* >dira+isa+iwa
diriwa TTT *tpt.* -ilwe. >dira+iwa
dirolola TTTT *tpt.* -itse. >dira+olola
dirololwa TTTT *tpt.* >dira+olola+iwa
dirwa TT *tpt.* >dira+iwa
dirwang TTG *tpt.* >dira+iwa+ng

omitted while less frequently used words such as ***direlang*** (1), ***dirisitswe*** (2)

dirisantse (1), ***dirileng*** (1), to mention but few, have been lemmatized. Consider Table 16 below:

Table 16: Overall frequency counts for the derivatives of the verb *dira* (7053) in WordSmith Tools

Rank	Word	Frequency		Rank	Word	Frequency
21286	dire	2,201		21454	dirisetswa	77
21426	dirisa	1,974		21167	diragale	76
21245	dirang	1,337		21136	dirafatsa	72
21401	dirile	1,098		21171	diragalelang	72
21322	direla	955		21170	diragalela	69
21473	dirisiwa	705		21302	diregile	58
21473	dirisiwa	705		21375	diretswe	52
21617	dirwa	654		21455	dirisetswang	51
21162	diragala	537		21479	dirisiwe	46
21163	diragalang	305		21300	direge	42
21476	dirisiwang	255		21334	direleng	33
21625	dirwang	253		21466	dirisitseng	33
21413	dirilwe	250		21376	diretsweng	30
21328	direle	229		21338	direlwe	29
21441	dirise	224		21303	diregileng	26
21193	diragetse	211		21157	diragadiwa	25
21414	dirilweng	185		21337	direlwang	23
21626	dirwe	181		21499	diriwang	23
21433	dirisang	174		21125	dirafalela	22
21190	diragatsa	164		21145	dirafetse	17
21297	direga	156		21173	diragalela	9
21470	dirisitswe	154		21182	diragaletseng	5
21371	diretse	146		21443	diriseditsweng	2
21464	dirisitse	143		21402	dirileng	1
21372	diretseng	107		21322	direlang	1
21472	dirisitsweng	105		21435	dirisantse	1
21340	direng	104				
21336	direlwa	101				
21119	dirafala	81				

It is noticed from the above examples that Kgasa and Tsonope (1995) apparently lemmatized these derivations without considering frequency of use offered by a large corpus, although they indicate that a corpus was consulted. For this reason, the dictionary does not address the needs of the users since derivations likely to be looked up by users are left out. Ideally one needs the overall frequencies as well as information on the spreading of these words across the different sources for the compilation of the lemmalist.

4.2.1 Physical limitation on the volume

According to Prinsloo (1994:94), limitations mostly on the number of pages or amount of entries that can be accommodated in a specific dictionary or sub-dictionary, has a far greater impact on the lemmatization of African languages than one would expect. Busane (1990:30) states that:

> "One of the basic problems of lexicography is to decide what to put in the dictionary and what to exclude"
> "…the problem remains as to whether all the lexical units that are likely to be derived from the main entry or the stem should be entered in the dictionary"

This implies that the need to select what to include or exclude in a dictionary proved to be a major concern for the compilers of the Setswana dictionaries.

Taking into consideration the limitations on the volume of Setswana dictionaries, the THAN contains 330 pages with approximately 16500 entries. This dictionary contains a huge number of derived forms presented as lemmas with elaborate comments on the form or morphological information without semantic information as reflected in example 4 above. This can lead to frustration and uncertainty about the exact meaning of the derivations and it is time consuming to try to find their meanings by artificially adding on the meaning of each derivation. (See *dikgakollišano* in Table 23 below).

4.2.2 Imbalances regarding the alphabetical stretches

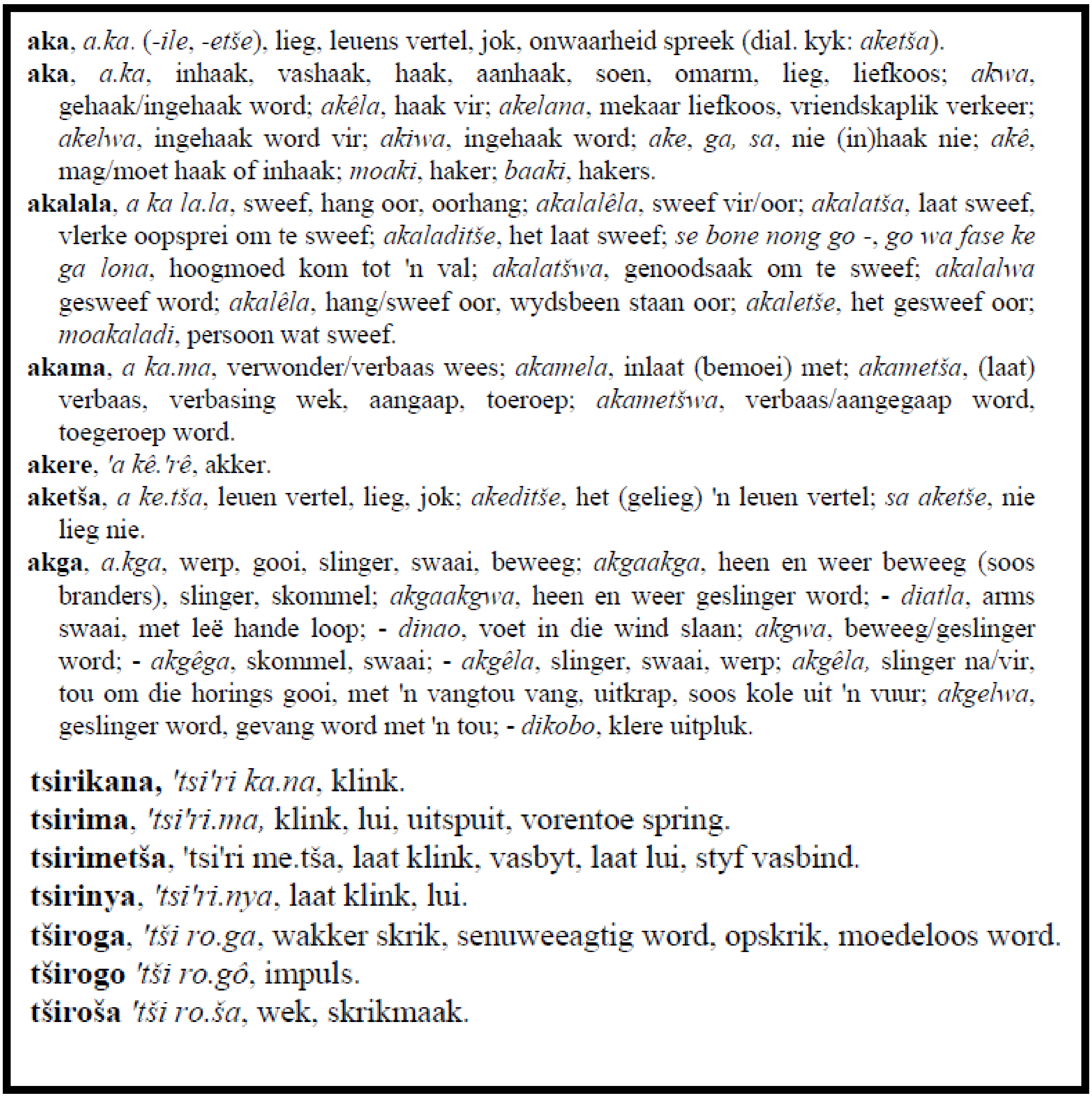
The second macrostructural problem apart from inclusion versus omission is the balancing out of the entire alphabetical categories of the Setswana dictionaries as a whole. This simply means that dictionary compilers often starts off treating the first few alphabetic categories exhaustively, but then grow tired towards the end of the alphabet. Compare (Prinsloo 2000a and Prinsloo 2000b) in reference to Kriel's dictionary. Landau 2001:398 states:

> "Nothing is more difficult to predict or control than a dictionary begun from scratch"

This remark is equally applicable to dictionaries that were compiled without the availability of a corpus. (See De Schryver and Prinsloo (2000) and Prinsloo and De Schryver (2003) for numerous examples of inconsistencies regarding over and under treatment in terms of alphabetical categories). Consider the following example below:

Example 5

PUKU1



aka, *a.ka.* (*-ile, -etše*), lieg, leuens vertel, jok, onwaarheid spreek (dial. kyk: *aketša*).

aka, *a.ka*, inhaak, vashaak, haak, aanhaak, soen, omarm, lieg, liefkoos; *akwa*, gehaak/ingehaak word; *akêla*, haak vir; *akelana*, mekaar liefkoos, vriendskaplik verkeer; *akelwa*, ingehaak word vir; *akiwa*, ingehaak word; *ake, ga, sa*, nie (in)haak nie; *akê*, mag/moet haak of inhaak; *moaki*, haker; *baaki*, hakers.

akalala, *a ka la.la*, sweef, hang oor, oorhang; *akalalêla*, sweef vir/oor; *akalatša*, laat sweef, vlerke oopsprei om te sweef; *akaladitše*, het laat sweef; *se bone nong go -, go wa fase ke ga lona*, hoogmoed kom tot 'n val; *akalatšwa*, genoodsaak om te sweef; *akalalwa* gesweef word; *akalêla*, hang/sweef oor, wydsbeen staan oor; *akaletše*, het gesweef oor; *moakaladi*, persoon wat sweef.

akama, *a ka.ma*, verwonder/verbaas wees; *akamela*, inlaat (bemoei) met; *akametša*, (laat) verbaas, verbasing wek, aangaap, toeroep; *akametšwa*, verbaas/aangegaap word, toegeroep word.

akere, *'a kê.'rê*, akker.

aketša, *a ke.tša*, leuen vertel, lieg, jok; *akeditše*, het (gelieg) 'n leuen vertel; *sa aketše*, nie lieg nie.

akga, *a.kga*, werp, gooi, slinger, swaai, beweeg; *akgaakga*, heen en weer beweeg (soos branders), slinger, skommel; *akgaakgwa*, heen en weer geslinger word; **-** *diatla*, arms swaai, met leë hande loop; **-** *dinao*, voet in die wind slaan; *akgwa*, beweeg/geslinger word; **-** *akgêga*, skommel, swaai; **-** *akgêla*, slinger, swaai, werp; *akgêla,* slinger na/vir, tou om die horings gooi, met 'n vangtou vang, uitkrap, soos kole uit 'n vuur; *akgelwa*, geslinger word, gevang word met 'n tou; **-** *dikobo*, klere uitpluk.

tsirikana, *'tsi'ri ka.na*, klink.
tsirima, *'tsi'ri.ma,* klink, lui, uitspuit, vorentoe spring.
tsirimetša, 'tsi'ri me.tša, laat klink, vasbyt, laat lui, styf vasbind.
tsirinya, *'tsi'ri.nya*, laat klink, lui.
tširoga, *'tši ro.ga*, wakker skrik, senuweeagtig word, opskrik, moedeloos word.
tširogo *'tši ro.gô*, impuls.
tširoša *'tši ro.ša*, wek, skrikmaak.

It is clear from example 5 that the first alphabetical words like *aka – akga* have been exhaustively treated while words towards the end of the alphabet like *tsirikana – tširoša* have received less attention.

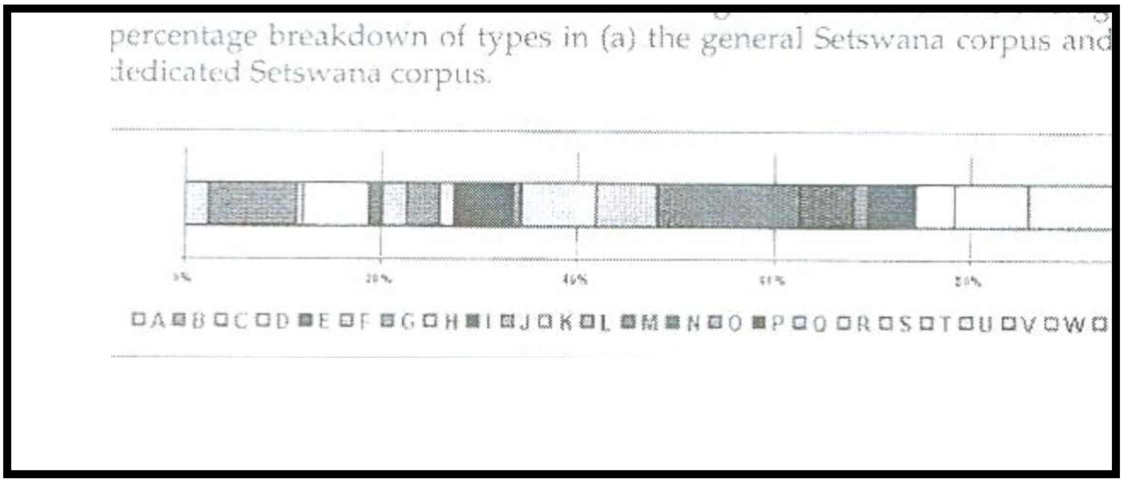
A multi-dimensional Setswana Ruler will now be introduced to study imbalances in Setswana dictionaries and to suggest a norm.

4.2.3 Building and applying a multi-dimensional Lexicographic Ruler

Prinsloo (2004) defines a Ruler as a practical instrument for measurement of the relative length of alphabetical stretches in alphabetically ordered dictionaries. Rulers are designed according to the generally accepted principle that alphabetical categories in any given language do not contain an equal number of words. Rulers are based upon the percentages of types per alphabetical category in corpora.

According to Prinsloo (2004:9) the real value of the Ruler lies in the fact that it focuses the attention of the compiler on potential ill-balanced areas, therefore the aim of the multi-dimensional Lexicographic Ruler for Setswana should be to eliminate the imbalances as reflected in Table 17 and Table 18 below.

Consider the Ruler for Setswana in Figure 23, based on the average of the percentage breakdown of types in a Setswana corpus.



percentage breakdown of types in (a) the general Setswana corpus and
dedicated Setswana corpus.

(Prinsloo, 2004:8)

Figure 24: A Ruler for Setswana

This ruler can also be expressed in terms of a percentage breakdown, i.e. divided into 100 blocks as a so-called block system.

Table 17: A block system for Setswana

1.	ALAF		21.	FELE		41.	KOUS		61.	MOTL		81.	SELE
2.	AROG		22.	FOLO		42.	LAEL		62.	MPHE		82.	SERA
3.	BADI		23.	GAGW		43.	LEBO		63.	NATE		83.	SETO
4.	BANN		24.	GATS		44.	LEKI		64.	NGWA		84.	SIMO
5.	BATW		25.	GOLO		45.	LERI		65.	NKUK		85.	SUAS
6.	BIRO		26.	GWET		46.	LETS		66.	NTEM		86.	TALE
7.	BOGA		27.	HUBE		47.	LOKO		67.	NTSH		87.	THAA
8.	BOLA		28.	IJES		48.	MAAD		68.	NYOR		88.	THIB
9.	BONK		29.	IKGO		49.	MAHA		69.	OOMA		89.	THWE
10.	BORU		30.	INOL		50.	MALE		70.	PANT		90.	TLAM
11.	BOUT		31.	IPUS		51.	MARA		71.	PHAK		91.	TLHA
12.	DAAM		32.	ITIS		52.	MATL		72.	PHIM		92.	TLHO
13.	DIFA		33.	ITSH		53.	MEFA		73.	PITL		93.	TLWA
14.	DIKG		34.	JOKO		54.	MESU		74.	PUDU		94.	TSAP
15.	DINK		35.	KANY		55.	MMAL		75.	RAMO		95.	TSHE
16.	DIRA		36.	KERO		56.	MMOL		76.	RENG		96.	TSHW
17.	DITH		37.	KGAR		57.	MOFI		77.	ROKG		97.	TSUN
18.	DITU		38.	KGOM		58.	MOKG		78.	RURU		98.	UBAU
19.	EGEP		39.	KHAN		59.	MONG		79.	SEBA		99.	WABO
20.	ETLH		40.	KODU		60.	MORW		80.	SEHI		100.	ZIMB

According to Prinsloo (2004:8), the Block System opens the door to a number of very practical applications. It gives a clear guidance in terms of page allocation, average length of articles, progress in terms of time and even remuneration intervals for part-time compilers.

From the above given statistics, it means that for a dictionary like THAN which contains 330 pages. 3,3 pages should roughly correlate with each block/percentage block. Consider now a comparison between MSED and the Ruler in terms of lemma percentage and page percentage.

Table 18: MSED versus the Ruler in terms of page allocation and the number of lemmas

	MSED:	**MSED:**	**Setswana**	**MSED lemmas**	**MSED Page %**
	Lemmas %	**Pages %**	**Ruler**	**vs. the Ruler**	**vs. Ruler**
A	1.2	1.3	2.6	-1.4	-1.3
B	4.7	4.6	9	-4.3	-4.4
C	0	0	0.6	-0.6	-0.6
D	6	6.4	6.6	-0.6	-0.2
E	1.2	1.3	1.4	-0.2	-0.1
F	3.7	3.3	2.4	1.3	0.9
G	5.2	5.3	3.4	1.8	1.9
H	0.9	0.9	1.5	-0.6	-0.6
I	5.3	4.9	5.9	-0.6	-1
J	0.7	0.7	0.8	-0.1	-0.1
K	12.2	11.9	7.7	4.5	4.2
L	6.7	6.8	6.1	0.6	0.7
M	12.5	13.7	14.6	-2.1	-0.9
N	4	4	5.5	-1.5	-1.5
O	1.3	1.3	1.6	-0.3	-0.3
P	5.9	6	4.6	1.3	1.4
Q	0	0.2	0.2	-0.2	0
R	3.9	3.5	3.9	0	-0.4
S	8.5	8.6	7.5	1	1.1
T	15.4	14.1	12.2	3.2	1.9
U	0.5	0.4	0.6	-0.1	-0.2
V	0	0	0.3	-0.3	-0.3
W	0.1	0.2	0.4	-0.3	-0.2
X	0	0.2	0.1	-0.1	0.1
Y	0.1	0	0.3	-0.2	-0.3
Z	0	0	0.2	-0.2	-0.2
		99.8	100		-0.2

The two Setswana dictionaries, i.e. THAN and THAND will now be compared in terms of pages utilized per alphabetical stretch and the number of the lemmas respectively.

Table 19: Comparison of the alphabetical stretch in THAN and THAND

(THAN) Tsonope (330 pages)			**(THAND) Kgasa and Tsonope** (126 pages)		
Category	Number of pages	Percentage	Category	Number of pages	Percentage
A	8	2,4	A	2	1,6
B	25	7,58	B	13	10,32
C	1	0,30	C	-	-
D	4	1,21	D	1	0,79
E	2	0,61	E	1	0,79
F	7	2,12	F	1	0,79
G	10	3,0	G	5	3,97
H	1	0,30	H	1	0,79
I	14	4,24	I	1	0,79
J	1	0,30	J	1	0,79
K	29	8,79	K	10	7,94
L	31	9,39	L	10	7,94
M	43	13,03	M	19	19,09
N	13	3,94	N	4	4,8
O	3	0,91	O	1	0,79
P	19	5,8	P	7	5,56
Q	1	00,30	Q	-	-
R	9	2,73	R	4	3,8
S	31	9,38	S	14	11,11
T	56	16,97	T	19	15,07
U	1	0,30	U	1	0,79
V	-		V	1	0,79
W	1	0,30	W	1	0,79
X	1	0,30	X	-	0,79
Y	1	0,30	Y	1	0,79
Z	-		Z	-	-

Table 19 indicates that most Setswana dictionaries have an alphabetical stretch of many pages for the lemmas B, K, L, M, P, S, and T. For a dictionary like the THAN, the stretches M and T fill a high number of pages 13, 03% and 16, 97% respectively. The same situation prevails for THAND with M and T at 19, 09% and 15, 07% respectively.

The difference in size for the alphabetical stretch M between these two dictionaries namely THAN and THAND could have been caused by the lexicographer's addition of new words and the ignorance of frequency counts, thus not taking a holistic approach. The alphabetical stretches for B, K, L, M, S and T in THAN occupy 65, 15% of the total number of 330 pages, and 67, 4% of the 126 number of pages for THAND.

Table 19 will now be graphically represented as indicated in Figures 25 and 26. The vertical axis indicates number of pages, the horizontal axis indicates category of the alphabetical stretches.

4.2.4 Graphical representation of the macrostructure of the two monolingual Setswana dictionaries

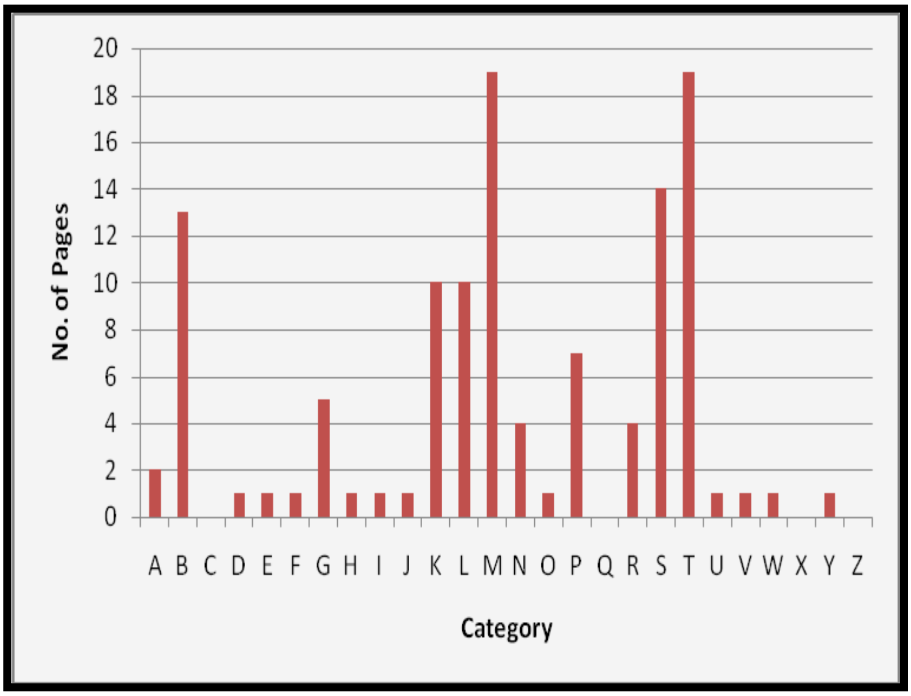


Figure 25: Graphical representation of THAND

It is clear from figure 25 that the alphabetical stretches for B, M, S, and T are relatively big and in particular contains large numbers of lemmas. The alphabetical stretches B, M and S contain the plural class prefix *ba*-, *me*- and *ma*- and the singular class prefix *se*- while the alphabetical stretches for T contains the majority of the Setswana verbs and few nouns.

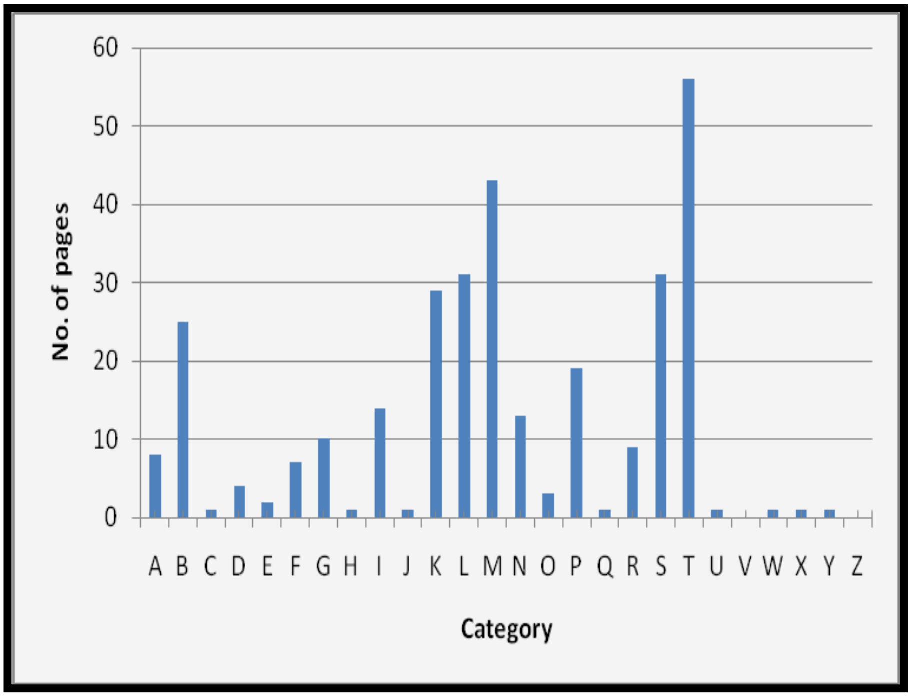


Figure 26: Graphical representation of THAN

Figure 26 contains the alphabetical stretches for M and T which are relatively big. The alphabetical stretch M and T contain the singular class prefixes *mo*-, and the plural class prefixes *ma*- and *me*- while the alphabetical stretches for T contains the majority of the Setswana verbs and nouns.

It is clear from both dictionaries that both the pages and the number for the lemma signs B, K, L, M, S and T occupy a huge number of pages and deserves exhausted treatment and E, H, J, Q, W, X and Y occupy a smaller number of pages while V and Z in THAN and C, Q and Z in THAND are empty, because the Setswana language does not contain words which start with V and Z.

4.3 Lemmatisation approaches, strategies and traditions

According to Prinsloo and Gouws (2005:85), it is important for the lexicographer when dealing with lemmatization in African languages to negotiate a complex interplay and overlap between (a) lemmatization approaches, (b) lemmatization strategies, (c) lexicographic traditions, (d) nominal and verbal structures and (e) conjunctiveness versus disjunctiveness. Compare the most relevant relations categorically in terms of columns A-E and rows 1-5.

Table 20: Lemmatization approaches, strategies, traditions, etc.

	A	**B**	**C**	**D**	**E**
	Lemmatization approaches	**Lemmatization strategies**	**Lexicographic traditions**	**Nominal and verbal structures**	**Conjuctiveness versus disjuctiveness**
1	Traditional	Stem	Word	Verbal prefixes	Conjunctive orthography
2	Paradigms	Singular and plural	stem	Verbal suffixes	Disjunctive orthography
3	Rule orientated	Singular only			
4	frequency	Left-expanded			
5		First or third letter			

In terms of Table 20 a complex set of 1-1 relations as given in Table 21 exists and has to be negotiated in any discussion of the lemmatization of nouns and verbs in African languages.

Table 21: Complex set of 1-1 relations in the Lemmatization's of nouns and verbs in African languages

A1:B1 ; A1:C1 ; A1:D1 ; A1:E1 ; A1:B2 ; A1:C2 ; A1:D2 ; A1:E2 ; A1:B3 ; A1:D3 ; A2:B1 ;
A2:C1 ; A2:D1 ; A2:E1 ; A2:B2 ; A2:C2 ; A2:D2 ; A2:E2 ; A2:B3 ; A2:D3 ; A3:B1 ; A3:C1 ;
A3:D1 ; A3:E1 ; A3:B2 ; A3:C2 ; A3:D2 ; A3:E2 ; A3:B3 ; A3:D3 ; A4:B1 ; A4:C1 ; A4:D1 ;
A4:E1 ; A4:B2 ; A4:C2 ; A4:D2 ; A4:E2 ; A4:B3 ; A4:D3 ; B1:C1 ; B1:D1 ; B1:E1 ; B1:C2 ;
B1:D2 ; B1:E2 ; B1:D3 ; B2:C1 ; B2:D1 ; B2:E1 ; B2:C2 ; B2:D2 ; B2:E2 ; B2:D3 ; B3:C1 ;
B3:D1 ; B3:E1 ; B3:C2 ; B3:D2 ; B3:E2 ; B3:D3 ; B4:C1 ; B4:D1 ; B4:E1 ; B4:C2 ; B4:D2 ;
B4:E2 ; B4:D3 ; B5:C1 ; B5:D1 ; B5:E1 ; B5:C2 ; B5:D2 ; B5:E2 ; B5:D3 ; C1:D1 ; C1:E1 ;
C1:D2 ; C1:E2 ; C1:D3 ; C2:D1 ; C2:E1 ; C2:D2 ; C2:E2 ; C2:D3 ; D1:E1 ; D1:E2 ; D2:E1 ;
D2:E2 ; D3:E1 ; D3:E2 ;

A brief overview of verbs will now be given to serve as a basis of the interpretation of the different lemmatisation strategies, approaches and traditions, as given in Table 21.

In the case of verbs numerous derivations of a single verb stem exist, consisting of the root plus one or more prefix(es) and or suffix(es) as is clearly indicated in Table 22 for the verb stem *reka* 'buy' which is structurally analysed in terms of 18 numbers.

Table 22: Derivative of reka (buy)

01	Root + standard modifications	VR	*reka*
		VRPre	*rekile*
		VRPas	*rekwa*
		VRPerPas	*rekilwe*
02 ANA	Root + reciprocal + standard modifications	VRRec	*rekana*
		VRRecPer	*rekane*
		VRRecPas	*rekangwa*
		VRRecPas	*rekangwe*
03 ANTSHA	Root + reciprocal + causative + standard modifications	VRRecCau	*rekantsha*
		VRRecCauPer	*rekantshitse*
		VRRecCauPas	*rekantshwa*
		VRRecCauPerPas	*rekantshitswe*
04 ANYA	root + alt. causative + standard modifications	VRAlt-Cau	*rekanya*
		VRAlt-CauPer	*rekantsha*
		VRAlt-CauPas	*rekangwa*
		VRAlt-CauPerPas	*rekangwe*
05 EGA	Root + neutron passive + standard modifications	VRNeu-Pas	*rekega*
		VRNeu-PasPer	*rekegile*
		VRPas	
		VRPerPas	
06 ELA	Root + applicative + standard modifications	VRApp	*rekela*
		VRAppPer	

		VRAppPas	
		VRAppPerPas	*reketswe*
07 ELANA	Root + applicative + reciprocal + standard modifications	VRAppRec	*rekelana*
		VRAppRecPer	*rekelane*
		VRAppRecPas	*rekelangwa*
		VRAppRecPerPas	*rekelangwe*
08 ISA	Root + causative + standard modifications	VRCau	*rekisa*
		VRCauPer	*rekisitse*
		VRCauPas	*rekisiwa*
		VRCauPerPas	*rekisitswe*
09 ISANA	Root + causative + reciprocal +standard modifications	VRCauRec	*rekisana*
		VRCauRecPer	*rekisane*
		VRCauRecPas	*rekisangwa*
		VRCauRecPerPas	*rekisangwe*
10 ISEGA	Root + causative + neutron passive + standard modifications	VRCauNpas	*rekisega*
		VRCauNpasPer	*rekisegile*
11 ISETSA	Root + causative + applicative + standard modifications	VRCappu	*rekisetsa*
		VRCauAppPer	*rekiseditse*
		VRCauAppPas	*rekisetswa*
		VRCauAppPerPas	*rekiseditswe*
12 ISETSANA	Root + causative + applicative + reciprocal + standard modifications	VRCauAppRec	*rekisetsana*
		VRCauAppRecPer	*rekisetsane*
		VRCauAppRecPas	*rekisetsanwa*
		VRCauAppRecPer Pas	*rekisetsanwe*

13 OLOLA	Root + reversive transitive + reciprocal + standard modification	VRRevt	*rekolola*
		VRRevtPer	*rekolotse*
		VRRevtPas	*rekololwa*
		VRRevtPerPas	*rekolotswe*
14 OLOLANA	Root + reversive transitive + reciprocal + standard modifications	VRRevtApp	*rekololana*
		VRRevtRecPer	*rekololane*
		VRRevtRecPas	*rekololanwa*
		VRRevtRecPerPas	*rekololanwe*
15 OLOLELA	Root + reversive transitive + applicative + standard modifications	VRRevtApp	*rekololela*
		VRRevtAppPer	*rekololetse*
		VRRevtAppPas	*rekololelwa*
		VRRetAppPerPas	*rekololetswe*
16 OLOLELANA	Root + reversive transitive + applicative + reciprocal + standard modifications	VRRevtApp	*rekololelana*
		VRRevtAppPer	*rekololelane*
		VRRevtAppPas	*rekololelanwa*
		VRRevtAppPerPas	*rekololelanwe*
17 OLODISA	Root + reversive transitive + causative + standard modifications	VRRevtCau	*rekolodisa*
		VRRevtCauPer	*rekolodisitse*
		VRRevtCauPerPas	*rekolodiswa*
18 OLODISANA	Root + reversive transitive + causative + reciprocal + standard modifications	VRRevtCauRec	*rekolodisana*
		VRRevtCauRecPer	*rekolodisane*
		VRRevtCauRecPas	*rekolodisangwa*

		VRRevtCauRecPas	*rekolodisangwe*

From Table 22, it is clear that the lexicographer has to consider a huge number of derivations per verb.

4.3.1 Lemmatization approaches

4.3.1.1 Traditional approach

This approach highlights a scenario whereby dictionary compilers seem to be unaware of the need to reduce the number of entries for a specific verb. Prinsloo (2004) defines the traditional approach as the worst situation where dictionary compilers fail to employ relevant or appropriate selection strategies and are even unaware of the problem of what to include in and what to omit from the dictionary. In case of nouns and verbs the compilers would e.g. conveniently ignore the need to reduce the number of derivations, which resulted in the compiler randomly adding words to the dictionary until the publication deadline.

In the preface to the SEAD, the compilers honestly admit:

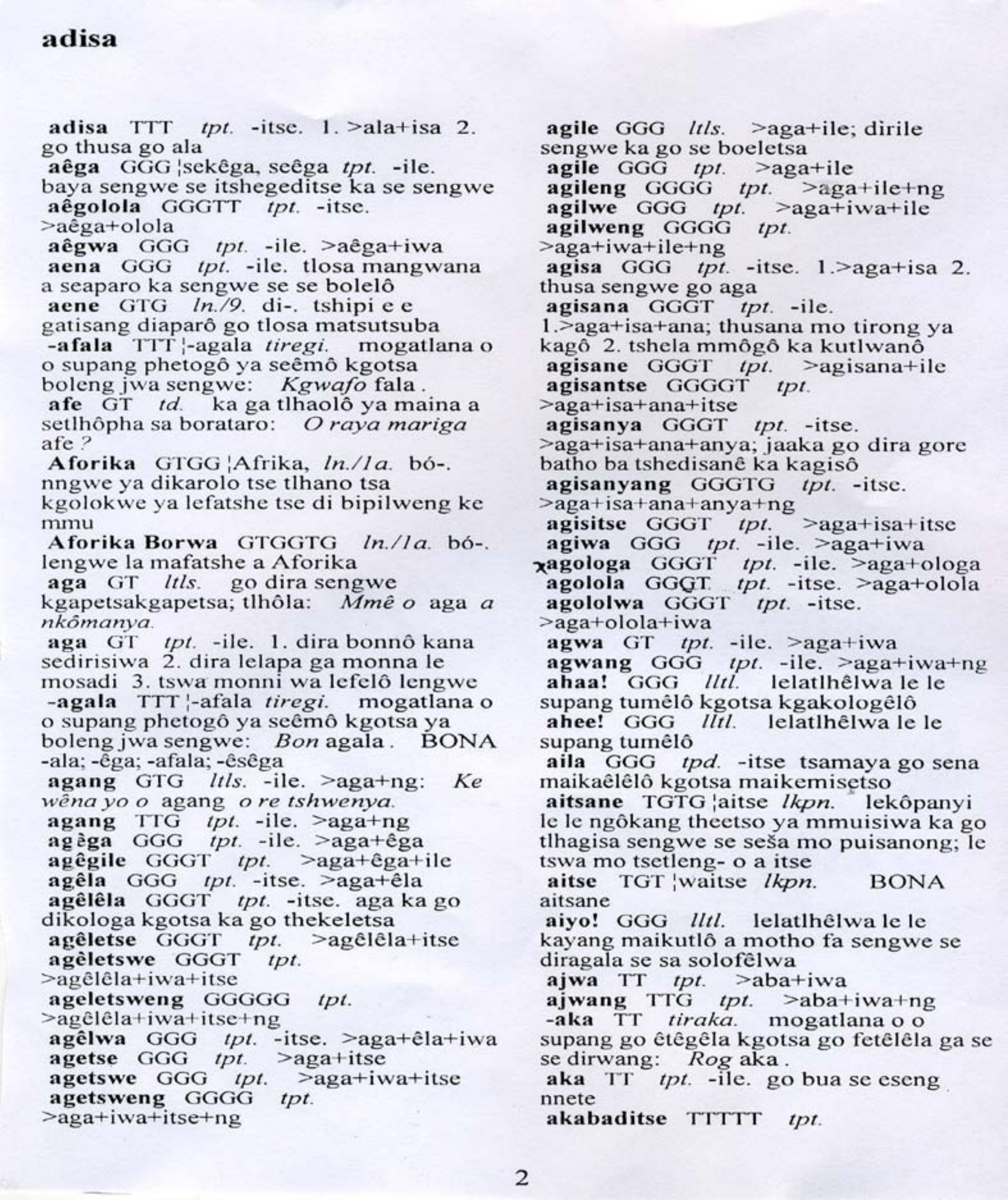
> "The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compilation of a dictionary. This can take place simply because the lexicographer had not encountered such words."

The traditional approach emphasizes the problem of essential words being accidentally excluded and a lot of rare words which are unlikely to be looked up by the target users are included. A typical example of this kind of an approach was discussed in Example 3 and Example 4 of this chapter where the imbalances arise from the traditional approach of the different derivations of verbs as highlighted. The issue whether frequency of use should be a determining factor for the inclusion versus omission of lemmas in dictionaries is often debated. It is relatively easy to define frequency as a guideline for inclusion/omission for bilingual dictionaries in the South African language lexicography context. Publishers normally limit the compiler to 5,000 lemmas in each side of the dictionary where target users include learners of the language. Given these requirements most of the 5,000 lemmas have to be selected on frequency of use. The situation becomes more problematic in monolingual dictionaries. If the monolingual dictionary is intended for mother-tongue speakers one could perhaps argue that they will not be looking for common frequently used words. Firstly, it is accepted that a mother-tongue speaker might not be inclined to look up the words such as *table* and *chair* for their meaning but for other purposes e.g. idiomatic use. Secondly monolingual dictionaries for Setswana include non-mother-tongue learners as target users who are likely to mainly look up frequently used words. Compilers of monolingual dictionaries are encouraged to compile rather comprehensive dictionaries, even as the first attempts for the language to include frequently used words but also a substantial number of infrequent words to solve the problem.

4.3.1.2 Completing paradigms

Prinsloo (1994:97) also calls this an ‘enter-them-all’ approach. In the THAN attempts were made to enter all nominal and verbal derivations to such an extent that mother-tongue speakers doubt whether many of these derivations are actually and actively used. Compare a section of the article of *aga* (build) in THAN in this regard:

Example 6: THAN



adisa

adisa TTT *tpt.* -itse. 1. >ala+isa 2. go thusa go ala

aêga GGG ¦sekêga, seêga *tpt.* -ile. baya sengwe se itshegeditse ka se sengwe

aêgolola GGGTT *tpt.* -itse. >aêga+olola

aêgwa GGG *tpt.* -ile. >aêga+iwa

aena GGG *tpt.* -ile. tlosa mangwana a seaparo ka sengwe se se bolelô

aene GTG *ln./9.* di-. tshipi e e gatisang diaparô go tlosa matsutsuba

-afala TTT ¦-agala *tiregi.* mogatlana o o supang phetogô ya seêmô kgotsa boleng jwa sengwe: *Kgwafo* fala .

afe GT *td.* ka ga tlhaolô ya maina a setlhôpha sa boratatro: *O raya mariga* afe ?

Aforika GTGG ¦Afrika, *ln./1a.* bó-. nngwe ya dikarolo tse tlhano tsa kgolokwe ya lefatshe tse di bipilweng ke mmu

Aforika Borwa GTGGTG *ln./1a.* bó-. lengwe la mafatshe a Aforika

aga GT *ltls.* go dira sengwe kgapetsakgapetsa; tlhôla: *Mmê o* aga *a nkômanya.*

aga GT *tpt.* -ile. 1. dira bonnô kana sedirisiwa 2. dira lelapa ga monna le mosadi 3. tswa monni wa lefelô lengwe

-agala TTT ¦-afala *tiregi.* mogatlana o o supang phetogô ya seêmô kgotsa ya boleng jwa sengwe: *Bon* agala . BONA -ala; -êga; -afala; -êsêga

agang GTG *ltls.* -ile. >aga+ng: *Ke wêna yo o* agang *o re tshwenya.*

agang TTG *tpt.* -ile. >aga+ng

agèga GGG *tpt.* -ile. >aga+êga

agêgile GGGT *tpt.* >aga+êga+ile

agêla GGG *tpt.* -itse. >aga+êla

agêlêla GGGT *tpt.* -itse. aga ka go dikologa kgotsa ka go thekeletsa

agêletse GGGT *tpt.* >agêlêla+itse

agèletswe GGGT *tpt.* >agêlêla+iwa+itse

ageletsweng GGGGG *tpt.* >agêlêla+iwa+itse+ng

agêlwa GGG *tpt.* -itse. >aga+êla+iwa

agetse GGG *tpt.* >aga+itse

agetswe GGG *tpt.* >aga+iwa+itse

agetsweng GGGG *tpt.* >aga+iwa+itse+ng

agile GGG *ltls.* >aga+ile; dirile sengwe ka go se boeletsa

agile GGG *tpt.* >aga+ile

agileng GGGG *tpt.* >aga+ile+ng

agilwe GGG *tpt.* >aga+iwa+ile

agilweng GGGG *tpt.* >aga+iwa+ile+ng

agisa GGG *tpt.* -itse. 1.>aga+isa 2. thusa sengwe go aga

agisana GGGT *tpt.* -ile. 1.>aga+isa+ana; thusana mo tirong ya kagô 2. tshela mmôgô ka kutlwanô

agisane GGGT *tpt.* >agisana+ile

agisantse GGGGT *tpt.* >aga+isa+ana+itse

agisanya GGGT *tpt.* -itse. >aga+isa+ana+anya; jaaka go dira gore batho ba tshedisanê ka kagisô

agisanyang GGGTG *tpt.* -itse. >aga+isa+ana+anya+ng

agisitse GGGT *tpt.* >aga+isa+itse

agiwa GGG *tpt.* -ile. >aga+iwa

x**agologa** GGGT *tpt.* -ile. >aga+ologa

agolola GGGT *tpt.* -itse. >aga+olola

agololwa GGGT *tpt.* -itse. >aga+olola+iwa

agwa GT *tpt.* -ile. >aga+iwa

agwang GGG *tpt.* -ile. >aga+iwa+ng

ahaa! GGG *lltl.* lelatlhêlwa le le supang tumêlô kgotsa kgakologêlô

ahee! GGG *lltl.* lelatlhêlwa le le supang tumêlô

aila GGG *tpd.* -itse tsamaya go sena maikaêlêlô kgotsa maikemisetso

aitsane TGTG ¦aitse *lkpn.* lekôpanyi le le ngôkang theetso ya mmuisiwa ka go tlhagisa sengwe se seša mo puisanong; le tswa mo tsetleng- o a itse

aitse TGT ¦waitse *lkpn.* BONA aitsane

aiyo! GGG *lltl.* lelatlhêlwa le le kayang maikutlô a motho fa sengwe se diragala se sa solofêlwa

ajwa TT *tpt.* >aba+iwa

ajwang TTG *tpt.* >aba+iwa+ng

-aka TT *tiraka.* mogatlana o o supang go êtêgêla kgotsa go fetêlêla ga se se dirwang: *Rog* aka .

aka TT *tpt.* -ile. go bua se eseng nnete

akabaditse TTTTT *tpt.*

2

It is clear from the above given example that users have to struggle through numerous columns of fine print to find the meaning of words such as *ageletsweng*, *agetsweng*, *agisanyang* etc. as illustrated above. The problem is clearly illustrated by Gouws and Prinsloo (2005:73). Consider the following example in terms of accessibility and unambiguous retrieval of the information from the perspective of an inexperienced learner of Sesotho sa Leboa. The user wants to look up the word *dikagollišano*. (S)he firstly has to strip the suffixes in order to find the verb stem and then to 'add' the semantic connotations in a cumulative way in order to find the meaning – thus up to 12 steps in total as given in Table 23 below.

Table 23: Accessibility and information retrieval process for *dikagollišano* in NSDN

1.	***dikagollišano***	↓	plural deverbative consisting of root + reversive transitive + causative + reciprocal + ending
2.	*kagollišano*	↓	singular deverbative consisting of root + reversive transitive + causative + reciprocal + ending
3.	*agollišana*	↓	verb root + reversive transitive + causative + reciprocal + ending
4.	*agolliša*	↓	verb root + reversive transitive + causative + ending
5.	agolla	↓	verb root + reversive transitive + ending
6.	**aga**	↓	**verb** (stem)
7.	**build**	↓	**meaning of the verb**
8.	break down	↓	reverse or opposite meaning 'un-build'
9.	cause to break down	↓	add causative sense of 'let/force'
10.	cause each other to break down	↓	add reciprocal sense of 'each other'
11.	the process of causing each other to break down	↓	change 'the process of …' to the plural
12.	**the process of causing each other to break down**		

(Gouws and Prinsloo 2005:40)

4.3.1.3 Rule-orientated approach

According to Prinsloo (1994:98), this approach presents a scenario where dictionary compilers still in principle, strive to complete paradigms and still attempt to make Provision for 'all' verbs and nouns and their derivatives. The compiler does not 'enter them all', but makes use of certain rules or guidelines which should be followed, if a word cannot be directly looked up in the dictionary. The target user is expected to interpret or reverse regularly derived derivations by means of a set of rules given in the non-alphabetical section of the dictionary as well as within the dictionary itself. In this regard the emphasis is on limiting the number of lemmas or sub lemmas in a dictionary. For example, (a) lemmatizing only the singular forms of nouns and only the basic verbal stems and (b) giving sets of rules in the user's guide for the user to strip suffixes and add meaning components. In principle, it still reflects the urge to 'enter-all' although it is quite economical in terms of dictionary space and it is also user-unfriendly.The strategy of lemmatizing singular forms of nouns as described in terms of THAND in 4.3.2 does however imply plural-to-singular guidance rules. A typical example is Pukuntšu (1989).

Example 7

PUKU 2 (Kriel and Van Wyk 1989: Preface)

Perfecta

-dile:	*-la*,	e.g.	*badile*	under	*bala*	
-ditše:	*-tša*,	e.g.	*biditše*	under	*bitša*	
-etše:	*-ela*,	e.g.	*rapetše*	under	*rapela*	
	-ala,	e.g.	*robeštše*	under	*robala*	
-tše:	*-as*,	e.g.	*bešite*	under	*beša*	
	-tšha,	e.g.	*bontšhitše*	under	*bontšha*	
	-sa,	e.g.	*lesitše*	under	*lesa*	
	-tswa,	e.g.	*hlatswitše*	under	*hlatšwa*	e.t.c

Applicative:

-etša:	*-ša*,	e.g.	*bešetsa*	under	*tloša*	
	-tšha,	e.g.	*tlhakantšhetša*	under	*ntšhetša*	
	-ša,	e.g.	*lesetša*	under	*leša*	
	-tšwa,	e.g.	*hlatswetša*	under	*hlatšwa*	
	-nya,	e.g.	*bofanya*		under	*bofa*
-letša:	*-tša*,	e.g.	*biletša*	under	*bitša*	

According to Prinsloo (2004:94), this approach on the other hand runs into serious difficulty with regard to practicality and user friendliness. Busane (1990:28) states its underlying disadvantages as follows:

> “---- many introductory pages are usually allocated to grammatical sketches of the language concerned without the knowledge of which is deemed hazardous to use the dictionary successfully. We believe, however, that these sections and introductory explanations are not sufficient provisions for a user friendly product. Dictionary users are known to allocate little time to the study of these prefatory matters.”

In this regard the compiler of a dictionary based on frequency of use can easily capitalise on the virtues of the rule-orientated approach.

4.3.1.4 Frequency-based approach

One could summarize the discussion thus far by stating that the corpus era for African languages introduced by Prinsloo (1991), opened new doors for the lemmatization of nouns and verbs namely lemmatization based on the frequency of use. According to Prinsloo (2004:93), using corpus data, the lexicographer can ensure that frequently used words are not accidentally omitted and, on the other hand, that precious dictionary space is not taken up by articles of lemmas which are unlikely to be looked up by the target users. Following this approach, the lexicographer can sensibly and drastically reduce the number of lemmas for a specific verb such as *rema* (chop) on frequency of use. The best point of departure is a frequency list of actual occurrences of the verb taken from a Setswana corpus.

Example 8

rema (34), remegang (5), reme (2), remege (1), remiremi (1), remile (35), remileng (77), remisa (1), remisang (100), remise (3), remiseditseng (2), remisetsa (1), remisetswe (1), remiseditsweng (4), remiso (1), remiswa (3), remiswang (4), remiswe (5).

The lexicographer can now lemmatize and reduce this list on the basis of the frequency of occurrence in the corpus. The lexicographer can for example decide not to lemmatize derivations containing the relative suffixes -*ng* and not to include infrequent derivations thus reducing the list to *rema, remile, remisa, remisitse, remetswe, remiswa*.

4.3.2 Lemmatization strategies

4.3.2.1 Lemmatization of nouns

According to Prinsloo (2004:95) lemmatization approaches are illustrated in a number of specific strategies such as lemmatizing: (a) both singular and plural, (b) only singular forms (c) noun stems, (d) on first or third letter and (e) using left expanded article structures.

Table 24: Noun classes and examples of Setswana

Class	**Prefix**	**Example**	**Translation**
1 2	*mo-* *ba-*	*monna* *banna*	man men
1a 2a	*Ø* *bo+*	*rrangwane* *borrangwane*	uncle uncles
3 4	*mo-* *me-*	*monwana* *menwana*	finger fingers
5 6	*le-* *ma-*	*lesogana* *masogana*	young man young men
7 8	*se-* *di-*	*selepe* *dilepe*	axe axes
9 10	*N-/Ø* *di+*	*nku* *dinku*	sheep sheep

11 12	*lo-* *di-*	*lobota* *dipota*	wall walls
13 14	*bo-* *ma-*	*bogobe* *magobe*	porridge different kinds of porridge
15	*go*	*go bona*	to see
16	*fa-*	*fase*	below
17	*go-*	*godimo*	above
18	*mo-*	*morago*	behind

(a) Lemmatizing both singular and plural noun forms

Prinsloo (2004:95) highlights that lemmatizing both singular and plural nouns is an extremely user-friendly strategy and very popular among inexperienced users and learners of a language. The user does not necessarily require previous knowledge. Unfortunately, the redundancy factor in terms of dictionary space is almost 80% high and has to be weighed up against the advantages in terms of the approach's user-friendliness and practicality. The compiler for MSED opted for lemmatising both singular and plural forms of nouns as suggested by Prinsloo and De Schryver (1999) and Gouws and Prinsloo (2005a:84-85). Compare example 9 in this regard.

Example 9: MSED

a	**b**
morara N. CL.3 *mo-*, SING. OF *merara,* ivy. **morutwa** N. CL.1 *mo-*, SING. OF *barutwa,* DER. F *rutwa,* same as *murutiwa* and *morutwi*, a Learner; disciple.	**merara** N. CL.4 *me-*, PL. OF *morara,* creeping plants; climbers, including Gymnema sylvestre and Secamone parvifolio; when used as a REl. s, it signifies branching in all directions. **barutwa** N. CL.2 *BA-*, PL. OF *morutwa,* students;

Note that in (9a and b) it is suggested that the treatment be given for the most frequent member of the singular form/plural pair. In the case of *morutwa* and *merara* full treatment is given while only basic information in *morara* and *barutwa* is given. This approach is in line with the more radical approach suggested by Gouws and Prinsloo (2005a), giving the treatment at the more frequently used member of the pair. For example, for *morara* and *morutwa* versus *merara* and *barutwa* in example 9, treatment is given to the singular form *morutwa* and the plural form *merara*. Consider example 10 where compilers of the SED opted to lemmatise both singular forms and plural forms giving equal treatment to both singular forms and plural forms.

Example 10: SED

a	**b**
Morutwa, N. CL.1 mo-, SING OF barutwa, DER. F. rutwa, same as morutiwa and morutwi, a learner; disciple. **Morara, n.,** A species of tree- climbing plant ; a wild vine.	**Barutwa,** N. CL.2 BA- PL. OF morutwa, students; scholars. **Merara,** n., pl. of morara, creeping plants; adj., Branching out in aal directions. *Ditshika tse di merara*. These veins spread or branch in all directions.

It is important to note in (10) that treatment is given to both members of the singular form/plural pair. Redundancy is increased to a great extend, although very user-friendly.

(b) Lemmatizing only singular noun forms

In terms of Prinsloo (2004:96), lemmatizing only singular noun forms, is a sound lexicographic strategy. It is not expecting too much from e.g. an advanced learner to know the regular productive rules of the language governing the formation of singular and plural forms, as illustrated in Table 25 which is an adaptation of the Sepedi rules for Setswana.

Table 25: Rules for looking up plural forms in Setswana

Rule		**Examples**	
word starts with	**look word up under**	**word start with**	**look word up under**
ba-	*mo-* *ngw-*	*basadi* *bana*	*mosadi* *ngwana*
bab-	*mm-*	*babetli*	*mmetli*
bo-	*(the stem)*	*bomalome*	*malome*
di-	*se-* *n-*	*dilepe* *dinku*	*selepe* *nku*
	(the stem) *lo-*	*dikgomo* *dinthe*	*kgomo* *lore*
ma-	*le-*	*maleme*	*leleme*
	bo-	*magobe*	*bogobe*
maj	*ma-*	*majang*	*bojang*
me-	*mo-* *me*	*megoma* *metsi*	*mogoma* *metsi*
meb-	*mm-*	*mebutla*	*mmutla*
mengw-	*ngw-*	*mengwaga*	*ngwaga*

A typical example of this kind of approach is found in THAN where the lemmatization strategy of singular nouns is followed. However inconsistencies regarding a number of words on the typical plural forms of the nouns *ba-* , *di-* and *me-* are noticed.

Although the editorial policy of dictionaries such as THAN is to lemmatize the singular forms only, plurals such as *badimo* (gods), *barwa* (sons), *baša* (youths) *and Bararo* (the three) are lemmatized. It is unfortunate that it is not always easy for users to look up plural forms under their singular forms because in most cases, from a user's point of view, there is no straightforward one to one correlation between some of the singular/plural class pairs first, *di–* (class 8 and class 10) has a one to two correlation. For example, the user who does not know the meaning of the word *dinku* (sheep) has to look it up under *senku or nku*). Secondly, a one to three correlation also exists in the case of *ma-* (class 6) and a one to four correlation in the case of *me–* (class 4). For example if the user is confronted with the word *metsi* (water). The rule in Table 25 states that *me*- should be looked up under *mo*- in class 4 and according to Table 25 it should be looked up under **motsi*, which is ungrammatical and the user has been misled in this regard**.** There are also irregular nouns that change their roots when taking the plural forms, e.g. singular *lore* (wood), plural *dinthe* (woods) and singular *lorako* (wall), plural *dithako* (walls) *in class 10.* Again the user who is not familiar with the language may get lost by simply adding the prefix (di-) * to the root -*rako* * *dirako* instead of (*dithako*) and *dinthe*.

(c) Lemmatizing nouns on the first or the third letter

Lemmatizing nouns on the first or the third letter is a method used by Snyman et al (1990) in their SEAD. Consider the following example:

Example 11: SEAD

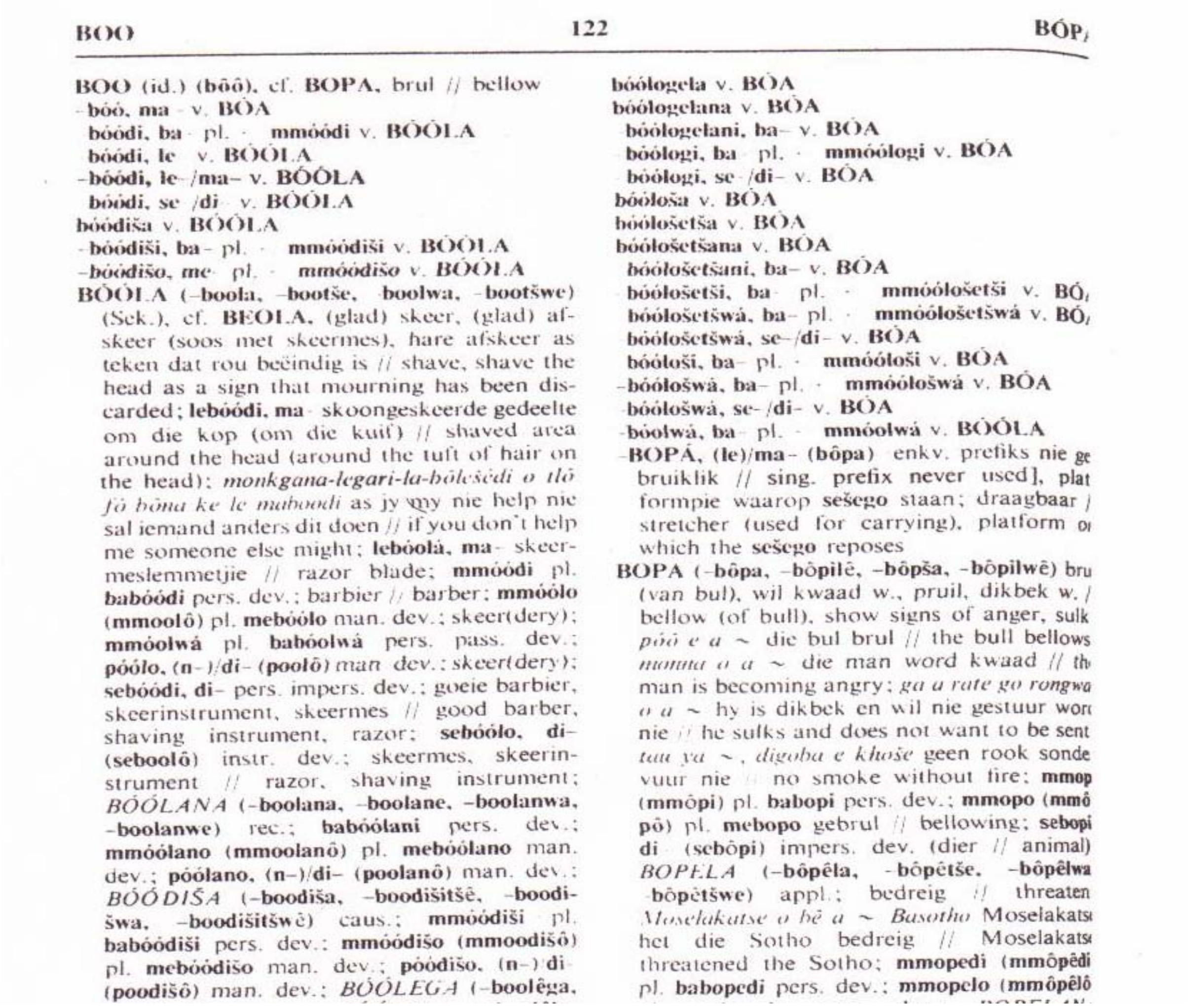
First letter	**Third letter**
kwáno (here)	*kwalo, le- ma-* (handwriting, orthography)
kwatla (a cut of meat from the back of a pig)	*kwapa, bo- le- ma-* (flat scale)
kwena (become strong and sturdy)	*kwé (kwê), le-* (Vaal river)
Laboraro (Wednesday)	*lapa, le- ma-* (traditional courtyard, home)

This approach has certain advantages for the inexperienced learner of Setswana, but can be frustrating to the user, because there are always two options to choose from when looking for the nouns. Redundancy is avoided by not having separate articles for lemmas such as *lekwalo* (letter) and *makwalo* (letters), etc.

(d) Lemmatizing noun stems

Lemmatizing nouns on their stems is a choice between the two conflicting lexicographic traditions namely, word versus stem followed in the lemmatization of African languages. This kind of an approach is not found in Setswana dictionaries. Consider the following examples of stem lemmatization taken from the CNSD and the SZD:

Example 12: CNSD

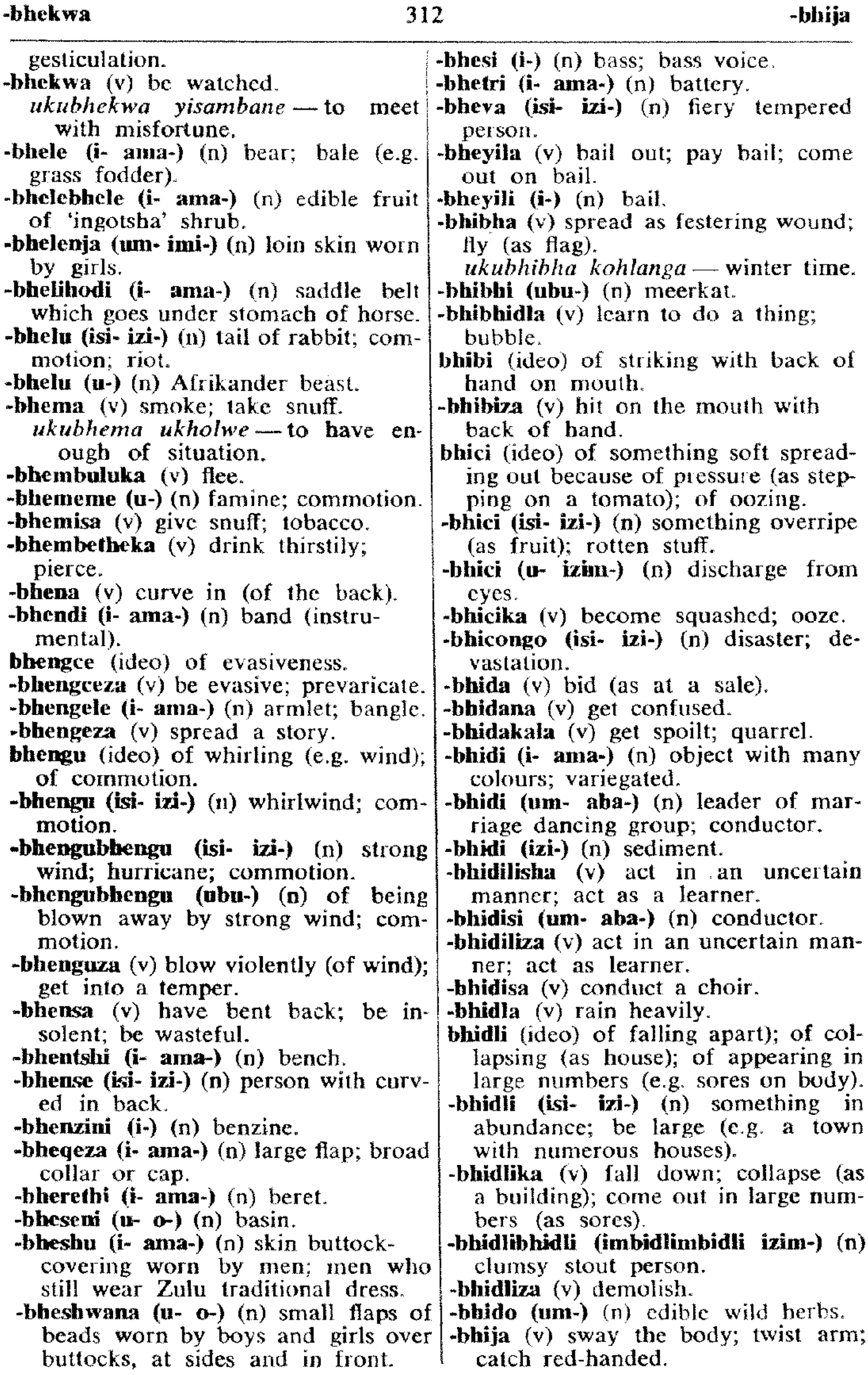


BOO 122 BÓP

BOO (id.) **(bôô)**, cf. **BOPA**, brul // bellow
-**bóó, ma**- v. **BÒA**
-**bóódi, ba**- pl. - **mmóódi** v. **BÒÓLA**
-**bóódi, le**- v. **BÒÓLA**
-**bóódi, le**-/**ma**- v. **BÓÓLA**
-**bóódi, se**-/**di**- v. **BÒÓLA**
bóódiša v. **BÒÓLA**
-**bóódiši, ba**- pl. - **mmóódiši** v. **BÒÓLA**
-***bóódišo, me***- pl. - ***mmóódišo*** v. **BÒÓLA**
BÒÓLA (-**boola**, -**bootše**, -**boolwa**, -**bootšwe**) (Sek.), cf. **BEOLA**, (glad) skeer, (glad) afskeer (soos met skeermes), hare afskeer as teken dat rou beëindig is // shave, shave the head as a sign that mourning has been discarded; **lebóódi, ma**- skoongeskeerde gedeelte om die kop (om die kuif) // shaved area around the head (around the tuft of hair on the head); *monkgana-legari-la-bôlešêdi o tlô fô bôna ke le maboodi* as jy my nie help nie sal iemand anders dit doen // if you don't help me someone else might; **lebóolá, ma**- skeermeslemmetjie // razor blade; **mmóódi** pl. **babóódi** pers. dev.; barbier // barber; **mmóólo (mmoolô)** pl. **mebóólo** man. dev.; skeer(dery); **mmóolwá** pl. **babóolwá** pers. pass. dev.; **póólo, (n-)/di- (poolô)** *man* dev.; *skeer(dery)*; **sebóódi, di**- pers. impers. dev.; goeie barbier, skeerinstrument, skeermes // good barber, shaving instrument, razor; **sebóólo, di- (seboolô)** instr. dev.; skeermes, skeerinstrument // razor, shaving instrument; *BÒÓLANA* (-**boolana**, -**boolane**, -**boolanwa**, -**boolanwe**) rec.; **babóólani** pers. dev.; **mmóólano (mmoolanô)** pl. **mebóólano** man. dev.; **póólano, (n-)/di- (poolanô)** man. dev.; *BÒÓDIŠA* (-**boodiša**, -**boodišitšê**, -**boodišwa**, -**boodišitšwê**) caus.; **mmóódiši** pl. **babóódiši** pers. dev.; **mmóódišo (mmoodišô)** pl. **mebóódišo** man. dev.; **póódišo, (n-)/di- (poodišô)** man. dev.; *BÒÓLEGA* (-**boolêga**,

bóólogela v. **BÒA**
bóólogelana v. **BÒA**
-**bóólogelani, ba**- v. **BÒA**
-**bóólogi, ba**- pl. - **mmóólogi** v. **BÒA**
-**bóólogi, se**-/**di**- v. **BÒA**
bóóloša v. **BÒA**
bóólošetša v. **BÒA**
bóólošetšana v. **BÒA**
-***bóólošetšani, ba***- v. **BÒA**
-**bóólošetši, ba**- pl. - **mmóólošetši** v. **BÒ**
-**bóólošetšwá, ba**- pl. - **mmóólošetšwá** v. **BÒ**
-**bóólošetšwá, se**-/**di**- v. **BÒA**
-**bóóloši, ba**- pl. - **mmóóloši** v. **BÒA**
-**bóólošwá, ba**- pl. - **mmóólošwá** v. **BÒA**
-**bóólošwá, se**-/**di**- v. **BÒA**
-**bóolwá, ba**- pl. - **mmóolwá** v. **BÒÓLA**
-**BOPÁ, (le)/ma- (bôpa)** enkv. prefiks nie gebruiklik // sing. prefix never used], platformpie waarop **sešego** staan; draagbaar / stretcher (used for carrying), platform on which the **sešego** reposes
BOPA (-**bôpa**, -**bôpilê**, -**bôpša**, -**bôpilwê**) bru (van bul), wil kwaad w., pruil, dikbek w. / bellow (of bull), show signs of anger, sulk *pôô e a ~* die bul brul // the bull bellows *monna o a ~* die man word kwaad // the man is becoming angry; *ga a rate go rongwa o a ~* hy is dikbek en wil nie gestuur word nie // he sulks and does not want to be sent *tau ya ~, digoba e khoše* geen rook sonder vuur nie // no smoke without fire; **mmop (mmôpi)** pl. **babopi** pers. dev.; **mmopo (mmôpô)** pl. **mebopo** gebrul // bellowing; **sebopi di (sebôpi)** impers. dev. (dier // animal) *BOPELA* (-**bôpêla**, -**bôpêtše**, -**bôpêlwa** -**bôpêtšwe**) appl.; bedreig // threaten *Moselakatse o bê a ~ Basotho* Moselakatse het die Sotho bedreig // Moselakatse threatened the Sotho; **mmopedi (mmôpêdi** pl. **babopedi** pers. dev.; **mmopelo (mmôpêlô**

xample 13: SZD



gesticulation.

-bhekwa (v) be watched.
ukubhekwa yisambane — to meet with misfortune.

-bhele (i- ama-) (n) bear; bale (e.g. grass fodder).

-bhelebhele (i- ama-) (n) edible fruit of 'ingotsha' shrub.

-bhelenja (um- imi-) (n) loin skin worn by girls.

-bhelihodi (i- ama-) (n) saddle belt which goes under stomach of horse.

-bhelu (isi- izi-) (n) tail of rabbit; commotion; riot.

-bhelu (u-) (n) Afrikander beast.

-bhema (v) smoke; take snuff.
ukubhema ukholwe — to have enough of situation.

-bhembuluka (v) flee.

-bhememe (u-) (n) famine; commotion.

-bhemisa (v) give snuff; tobacco.

-bhembetheka (v) drink thirstily; pierce.

-bhena (v) curve in (of the back).

-bhendi (i- ama-) (n) band (instrumental).

bhengce (ideo) of evasiveness.

-bhengceza (v) be evasive; prevaricate.

-bhengele (i- ama-) (n) armlet; bangle.

-bhengeza (v) spread a story.

bhengu (ideo) of whirling (e.g. wind); of commotion.

-bhengu (isi- izi-) (n) whirlwind; commotion.

-bhengubhengu (isi- izi-) (n) strong wind; hurricane; commotion.

-bhengubhengu (ubu-) (n) of being blown away by strong wind; commotion.

-bhenguza (v) blow violently (of wind); get into a temper.

-bhensa (v) have bent back; be insolent; be wasteful.

-bhentshi (i- ama-) (n) bench.

-bhense (isi- izi-) (n) person with curved in back.

-bhenzini (i-) (n) benzine.

-bheqeza (i- ama-) (n) large flap; broad collar or cap.

-bherethi (i- ama-) (n) beret.

-bheseni (u- o-) (n) basin.

-bheshu (i- ama-) (n) skin buttock-covering worn by men; men who still wear Zulu traditional dress.

-bheshwana (u- o-) (n) small flaps of beads worn by boys and girls over buttocks, at sides and in front.

-bhesi (i-) (n) bass; bass voice.

-bhetri (i- ama-) (n) battery.

-bheva (isi- izi-) (n) fiery tempered person.

-bheyila (v) bail out; pay bail; come out on bail.

-bheyili (i-) (n) bail.

-bhibha (v) spread as festering wound; fly (as flag).
ukubhibha kohlanga — winter time.

-bhibhi (ubu-) (n) meerkat.

-bhibhidla (v) learn to do a thing; bubble.

bhibi (ideo) of striking with back of hand on mouth.

-bhibiza (v) hit on the mouth with back of hand.

bhici (ideo) of something soft spreading out because of pressure (as stepping on a tomato); of oozing.

-bhici (isi- izi-) (n) something overripe (as fruit); rotten stuff.

-bhici (u- izim-) (n) discharge from eyes.

-bhicika (v) become squashed; ooze.

-bhicongo (isi- izi-) (n) disaster; devastation.

-bhida (v) bid (as at a sale).

-bhidana (v) get confused.

-bhidakala (v) get spoilt; quarrel.

-bhidi (i- ama-) (n) object with many colours; variegated.

-bhidi (um- aba-) (n) leader of marriage dancing group; conductor.

-bhidi (izi-) (n) sediment.

-bhidilisha (v) act in an uncertain manner; act as a learner.

-bhidisi (um- aba-) (n) conductor.

-bhidiliza (v) act in an uncertain manner; act as learner.

-bhidisa (v) conduct a choir.

-bhidla (v) rain heavily.

bhidli (ideo) of falling apart); of collapsing (as house); of appearing in large numbers (e.g. sores on body).

-bhidli (isi- izi-) (n) something in abundance; be large (e.g. a town with numerous houses).

-bhidlika (v) fall down; collapse (as a building); come out in large numbers (as sores).

-bhidlibhidli (imbidlimbidli izim-) (n) clumsy stout person.

-bhidliza (v) demolish.

-bhido (um-) (n) edible wild herbs.

-bhija (v) sway the body; twist arm; catch red-handed.

In (12) the Sepedi nouns *lebopa* and *mabopa* are lemmatized on their nominal stem form *–bopa.* The same holds true for this isiZulu nouns *ibhendi* and *amabhendi* lemmatised under *-bhendi*.

Lemmatising nouns stems is not user-friendly especially for inexperienced users and learners of the language and it introduces unnecessary problems in respect of stem identification. Central to these traditions stands the issue of conjunctivism versus disjunctivism. Consider an example where Setswana (disjuntivism) is compared with isiZulu (conjuctivism).

4.3.3 Conjunctiveness versus disjunctiveness

Table 26: Conjunctivism versus disjunctivism

Setswana		**isiZulu**	
Re a leboga	(We are thankful)	*Siyabonga*	(We are thankful)
Ke ne ke opela	(I was singing)	*Bengicula*	(I was singing)
A re kopaneng	(Let us meet)	*Masihlangane*	(Let us meet)

For words in case of ‘We are thankful’, ‘I was singing’ in Setswana are written as four separate words while in isiZulu are written as a single word. Van Wyk (1995) describes conjunctivism versus disjunctivism as purely a matter of orthographical convention. The stem tradition has mostly been followed for dictionaries for the conjunctively written languages namely isiZulu, isiXhosa, isiNdebele and Siswati (Nguni languages) and the word tradition for the disjunctively written ones, Sesotho sa Leboa, Setswana, Sesotho, Tshivenda and Xitsonga.

4.3.4 Lexicographic Traditions

4.3.4.1 Stem versus word tradition in respect of verbs

Van Wyk (1995) states that it is important to note the difference between nouns and verbs when it comes to affixes (prefixes and suffixes). According to Prinsloo (2009:6), lemmatising stems of verbs in particular makes sense for the conjunctively written languages. A huge number of prefixes, up to more than 4000 per verb, combine freely and productively with verbs, such as subject concords, object concords, negative morphemes, the progressive, the potential, future, etc. Lexicographers of conjunctively as well as disjunctively written languages agree that stem lemmatisation is the best option. It would also be totally redundant to lemmatise each verb stem plus prefixes separately. For example, *ngiyabonga* (I am thankful) and *masihlangane* (let us meet) etc. in isiZulu are all lemmatised under their stem -*bonga* (thankful) and *hlangane* (meet). According to Prinsloo (2009:9), in case of verbal suffixes in particular, verb stems plus suffixes should be lemmatised separately to avoid very long articles where treatment of the numerous derivations is attempted under a single stem form, for example, as in the Popular Northern Sotho Dictionary (POP) in contrast to the above.

Example 14: **POP**

badiša	cause to read/count
bala	read; count, reckon: include
balêga	be counted
balêgê go se	~ innumerable
balêla	read/count for…
balola	recount…
balwa	be read, counted, ~ ***le*** included

4.4 Problems regarding the lemmatization of nouns with prefixes *le*- and *lo*- in Setswana dictionaries

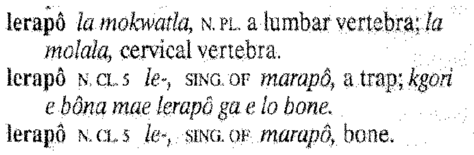
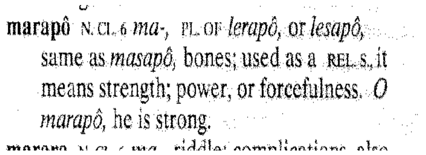
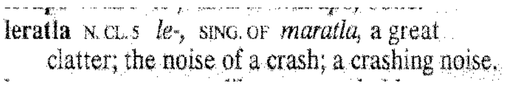
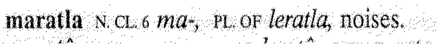
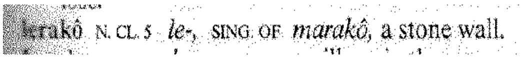
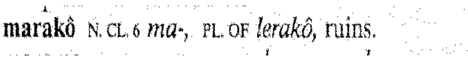
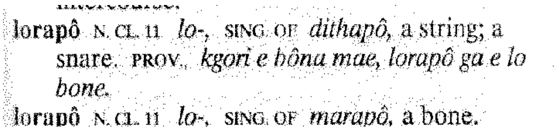
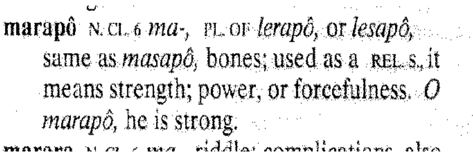
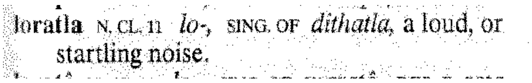
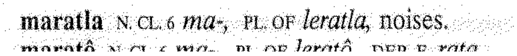
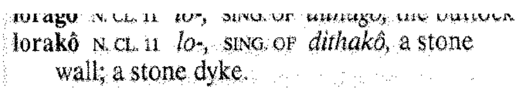
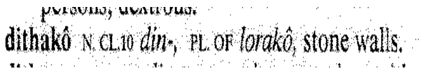
In Setswana, particularly the South African written variant, use is made of the singular class prefixes *lo-/le-* with their accompanying plural forms *di-/ma-* respectively. It has to be clarified that even in dictionary entries, *lo*-, *le*-, *di*- and *ma*- should be captured differently to avoid a situation where one excludes another. For example, in Setswana sometimes we speak of *lofuka* (wing) whose plural form is *diphuka* (wings), *lofofa* (wing) whose plural form is *diphofa* (wings) while in other instances we will use *lefofa* (wing) with the plural *mafofa* (wings). While the two versions are intelligible in all instances, *diphofa* includes references to big feathers and a multiplicity of feathers while *mafofa* simply refers to plurality. It has to be borne in mind that, while class prefixes *lo*- and *le*- can be used interchangeably in some instances, their usage becomes absolute in others. For example, there are instances where *lo*- cannot be replaced by *le*- as in *lore* (wood) whose plural is *dinthe* (woods), and similarly, *le*- cannot be substituted for *lo*- as in *lesepa* (mess) whose plural form is *masepa* (mess). Also in terms of concordial forms, *lo*- and *le*- used variably after *lona* as in '***lo**na **lo** bona **lo** le batho*' (you think you are people) and '***le**na **le** bona **le** le batho'*. (you think you are people). *Lo*- in this regard serves as a variant of *le*- although the two are used interchangeably. It is recommended that the written variant be used consistently when chosen, that is, if *lo*- is used, it has to be used in the whole text. The same applies to *le*-. Be that as it may, *lo*- is used consistently in some editions of the Bible and it has acquired a sense of religiosity and respect and accords these qualities to the addresses.

A number of nouns of class 5 *le*- are normally used in the plural only. The *lo*- class prefix belongs to the noun class 11 and takes the plural prefix *diN*- and it is typically found in the central and southern dialects of Setswana.

According to Cole (1975:91), this class has been partially or completely absorbed into class 5 *le*-. The singular form of class 6 has the singular prefix *le*-, but retains the plural prefix diN-, though frequently replaced by the prefix *ma*-. It is indicated that class 11 *lo*- is gradually becoming obsolete in Setswana.

From the above given examples, it is clear that there is no fixed law or principles of dealing with the lemmatization of the noun prefixes *le*- and *lo*-. Consider the following examples 15, 16, and 17 below:

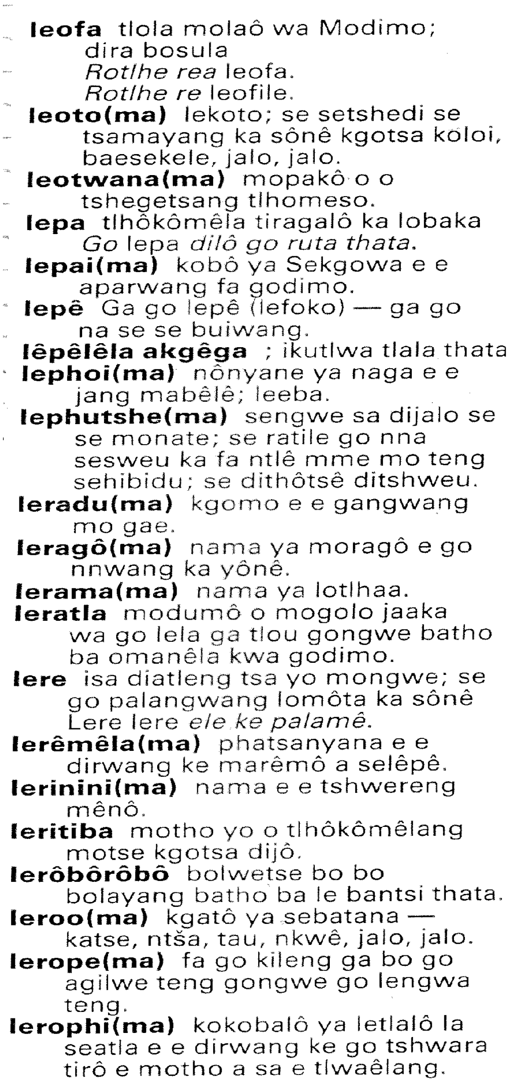
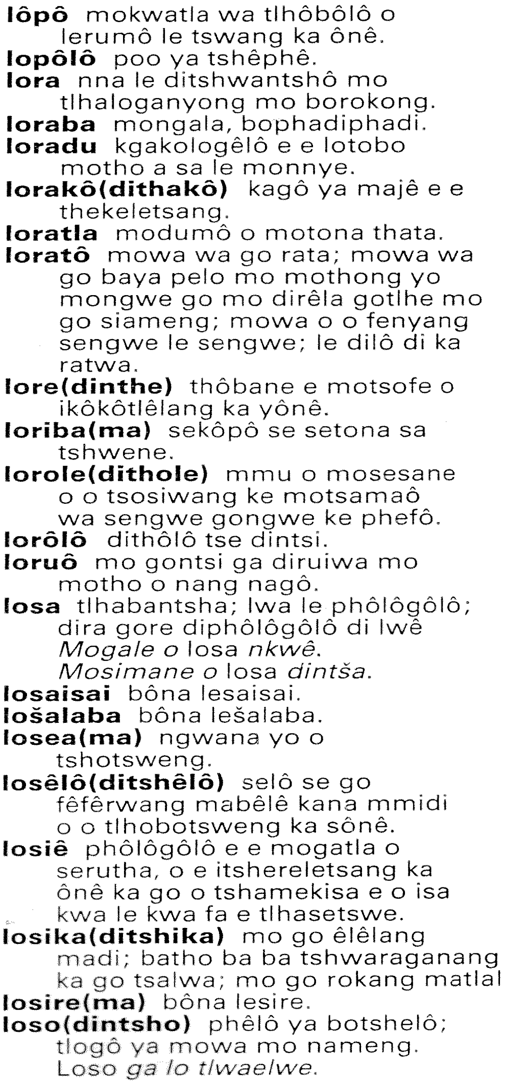
Example 15: MSED

(a) Singular *le-*	**Plural**
Lerapo  **lerapô** *la mokwatla,* N. PL. a lumbar vertebra; *la molala,* cervical vertebra. **lerapô** N. CL. 5 *le-,* SING. OF *marapô,* a trap; *kgori e bôna mae lerapô ga e lo bone.* **lerapô** N. CL. 5 *le-,* SING. OF *marapô,* bone.	***Marapo***  **marapô** N. CL. 6 *ma-,* PL. OF *lerapô,* or *lesapô,* same as *masapô,* bones; used as a REL. S., it means strength; power, or forcefulness. *O marapô,* he is strong.
Leratla  **leratla** N. CL. 5 *le-,* SING. OF *maratla,* a great clatter; the noise of a crash; a crashing noise.	***Maratla***  **maratla** N. CL. 6 *ma-,* PL. OF *leratla,* noises.
Lerako  **lerakô** N. CL. 5 *le-,* SING. OF *marakô,* a stone wall.	***Marako***  **marakô** N. CL. 6 *ma-,* PL. OF *lerakô,* ruins.
(b) Singular *lo-*	**Plural**
Lorapo  **lorapô** N. CL. 11 *lo-,* SING. OF *dithapô,* a string; a snare. PROV., *kgori e bôna mae, lorapô ga e lo bone.* **lorapô** N. CL. 11 *lo-,* SING. OF *marapô,* a bone.	***Marapo***  **marapô** N. CL. 6 *ma-,* PL. OF *lerapô,* or *lesapô,* same as *masapô,* bones; used as a REL. S., it means strength; power, or forcefulness. *O marapô,* he is strong.
Loratla  **loratla** N. CL. 11 *lo-,* SING. OF *dithatla,* a loud, or startling noise.	***Maratla***  **maratla** N. CL. 6 *ma-,* PL. OF *leratla,* noises.
Lorako  **lorakô** N. CL. 11 *lo-,* SING. OF *dithakô,* a stone wall; a stone dyke.	***Dithako***  **dithakô** N. CL.10 *din-,* PL. OF *lorakô,* stone walls.

The compilers of the MSED opted to lemmatise both the singular and the plural forms of the nouns of class prefixes *le*- and *lo*- separately and give treatment to both pairs. However, redundancy becomes a factor in dictionaries that offer treatment of both the singular and plural forms as indicated in example 16 above.

Consider another example where compilers opted to lemmatise only the most frequent forms of both the singular and the plural forms of the prefixes *le*- and *lo*.

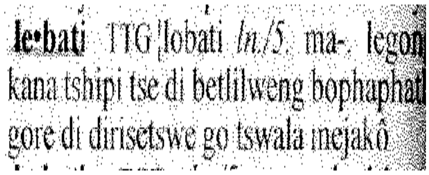
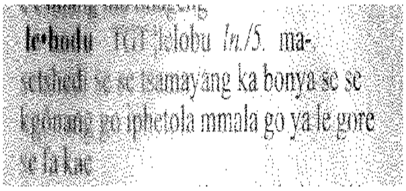
Example 16: THAND

Le-	***Lo-***
 **leofa** tlola molaô wa Modimo; dira bosula *Rotlhe rea* leofa. *Rotlhe re* leofile. **leoto(ma)** lekoto; se setshedi se tsamayang ka sônê kgotsa koloi, baesekele, jalo, jalo. **leotwana(ma)** mopakô o o tshegetsang tlhomeso. **lepa** tlhôkômêla tiragalô ka lobaka *Go* lepa *dilô go ruta thata*. **lepai(ma)** kobô ya Sekgowa e e aparwang fa godimo. **lepê** Ga go lepê (lefoko) — ga go na se se buiwang. **lêpêlêla akgêga** ; ikutlwa tlala thata **lephoi(ma)** nônyane ya naga e e jang mabêlê; leeba. **lephutshe(ma)** sengwe sa dijalo se se monate; se ratile go nna sesweu ka fa ntlê mme mo teng sehibidu; se dithôtsê ditshweu. **leradu(ma)** kgomo e e gangwang mo gae. **leragô(ma)** nama ya moragô e go nnwang ka yônê. **lerama(ma)** nama ya lotlhaa. **leratla** modumô o mogolo jaaka wa go lela ga tlou gongwe batho ba omanêla kwa godimo. **lere** isa diatleng tsa yo mongwe; se go palangwang lomôta ka sônê Lere lere *ele ke palamê*. **lerêmêla(ma)** phatsanyana e e dirwang ke marêmô a selêpê. **lerinini(ma)** nama e e tshwereng mênô. **leritiba** motho yo o tlhôkômêlang motse kgotsa dijô. **lerôbôrôbô** bolwetse bo bo bolayang batho ba le bantsi thata. **leroo(ma)** kgatô ya sebatana — katse, ntša, tau, nkwê, jalo, jalo. **lerope(ma)** fa go kileng ga bo go agilwe teng gongwe go lengwa teng. **lerophi(ma)** kokobalô ya letlalô la seatla e e dirwang ke go tshwara tirô e motho a sa e tlwaêlang.	 **lôpô** mokwatla wa tlhôbôlô o lerumô le tswang ka ônê. **lopôlô** poo ya tshêphê. **lora** nna le ditshwantshô mo tlhaloganyong mo borokong. **loraba** mongala, bophadiphadi. **loradu** kgakologêlô e e lotobo motho a sa le monnye. **lorakô(dithakô)** kagô ya majê e e thekeletsang. **loratla** modumô o motona thata. **loratô** mowa wa go rata; mowa wa go baya pelo mo mothong yo mongwe go mo dirêla gotlhe mo go siameng; mowa o o fenyang sengwe le sengwe; le dilô di ka ratwa. **lore(dinthe)** thôbane e motsofe o ikôkôtlêlang ka yônê. **loriba(ma)** sekôpô se setona sa tshwene. **lorole(dithole)** mmu o mosesane o o tsosiwang ke motsamaô wa sengwe gongwe ke phefô. **lorôlô** dithôlô tse dintsi. **loruô** mo gontsi ga diruiwa mo motho o nang nagô. **losa** tlhabantsha; lwa le phôlôgôlô; dira gore diphôlôgôlô di lwê *Mogale o* losa *nkwê*. *Mosimane o* losa *dintša*. **losaisai** bôna lesaisai. **lošalaba** bôna lešalaba. **losea(ma)** ngwana yo o tshotsweng. **losêlô(ditshêlô)** selô se go fêfêrwang mabêlê kana mmidi o o tlhobotsweng ka sônê. **losiê** phôlôgôlô e e mogatla o serutha, o e itshereletsang ka ônê ka go o tshamekisa e o isa kwa le kwa fa e tlhasetswe. **losika(ditshika)** mo go êlêlang madi; batho ba ba tshwaraganang ka go tsalwa; mo go rokang matlal **losire(ma)** bôna lesire. **loso(dintsho)** phêlô ya botshelô; tlogô ya mowa mo nameng. Loso *ga lo tlwaelwe*.

This approach reduces redundancy, but is heavily dependent on previous knowledge of the language.

Consider the third example taken from the THAN where the class prefixes *le*- and *lo*- are both lemmatised, and instead of treating them, they supplied a cross-reference to the prefix *le*-. Consider for example, nouns such as *lonao*, *longana*, *loologa*, *lookwane*, *looto*, to mention but few.

Example 17: THAN

Lebati	***Lobati***
 **le•bati** TTG lobati *ln./5.* ma- legon kana tshipi tse di betlilweng bophaphat gore di dirisetswe go tswala mejakô	
Lebodu	***Lobodu***
 **le•bodu** TGT lelobu *ln./5.* ma- setshedi se se tsamayang ka bonya se se kgônang go iphetola mmala go ya le gore se la kae	
Lebone	***Lobone***
Lebota	***Lobota***

This strategy increases redundancy to a great extent and a number of cross-references cause a great deal of page turning which is time consuming. A further complication arises from the fact that the relations between singular and plural are not always regular and not satisfactorily treated in dictionaries such as MSED and THAND. Consider the following example in this regard:

Example 18

Singular		**Plural**
Letsatsi (sun / day) *[Matsatsi]*	>	*Malatsi*
Letsogo; Lebogo (hand) *[Mabogo]*	>	*Matsogo*
Letsele; Lebele (breast) *[Matsele]*	>	*Mabele*

From the above example it is clear that preference is given *to malatsi*, *matsogo* and *mabele* that are irregular forms. The latter is further supported by an example of a proverb *mabogo dinku a a thebana* (it is important to assist one another).

The above mentioned nouns, i.e. *matsatsi* and *malatsi; matsogo* and *mabogo* etc. should be treated as variants and lemmatized as single entries. An example should also be given to help identify subtle differences between similar words as indicated in brackets *[malatsi]*, *[matsogo] and [matsele*] which are preferred to *matsatsi*, *mabogo* and *mabele.*

Furthermore, both the THAN and the THAND lemmatize singular nouns only and give the plural prefix as part of the treatment. This can be confusing to the user. Suppose a foreign user is confronted with the word *malatsi*. (S)he will look for the word under **lelatsi*. In this instance the user will get lost since the plural form has undergone a morphological change. It is recommended that both singular and plural forms of nouns be lemmatized to render a user-friendly product. This approach should also be based on a frequency count as outlined previously in this chapter.

4.5 The absence of nominal derivations

Dictionary information on nominal derivations such as diminutives are not provided by Setswana lexicographers. It is unfortunate that the majority of Setswana dictionaries provide inadequate treatment of these lexical items. It is interesting to note that diminutives can have different meanings (senses) when nouns referring to various body organs or human beings are used. Diminutives associated with body organs express behavioural patterns, i.e. gossip and idling around.

Example 19

(a) Diminitives associated with body organs

O tota o le loleme	> (You really have a tongue) 'you talk too much'
*O tota o le lole**ngwana***	> (You really have little tongue) 'you gossip too much'
O lenao	> (You are a foot) 'you are loitering'
*O lenao**nyana***	> (You are a little foot)'you are idling or gallivanting'

The suffixes -*ngwana* and -*nyana in lole**ngwana** and lenao**nyana*** justify treatment in their own right.

(b) Human beings

i. *Bona **mosadinyana** yoo*. (Look at that little woman).
ii. *Ao **mosadinyana** wa me*. (Oh my little woman) Oh my adorable woman.

In the above given example in b(i), the sentence can refer to the following:

- ❖ Look at that adorable woman
- ❖ Look at that little woman
- ❖ Look at that little woman showing disrespect but may also indicate an expression of love or passion

Example b (ii) indicates an expression of love or passion.

The above-mentioned assertion underlines the need to be included and treated in the dictionary. The absence of the oral corpus and the exclusion of the diminutive can deprive the user of certain information, especially if there is a high frequency of usage.

Consider example 20 below where diminutives have undergone morphological change:

Example 20

Word			**Diminutive - form**	
Tsebe	(ear)	>	*Tse**jwana***	(small ear)
Moriri	(hair)	>	*Mori**tshana***	(small hair)
Legodu	(thief)	>	*Lego**tswana***	(small thief)
Kgomo	(cow)	>	*Kgo**ngwana***	(calf)
Ngwana	(child)	>	*Ngwa**nyana***	(little baby)

From the above given examples it can be seen that the diminutives have undergone morphological change. The user who is not familiar with the language may easily be confused and will be unable to look them up under their non-derived forms.

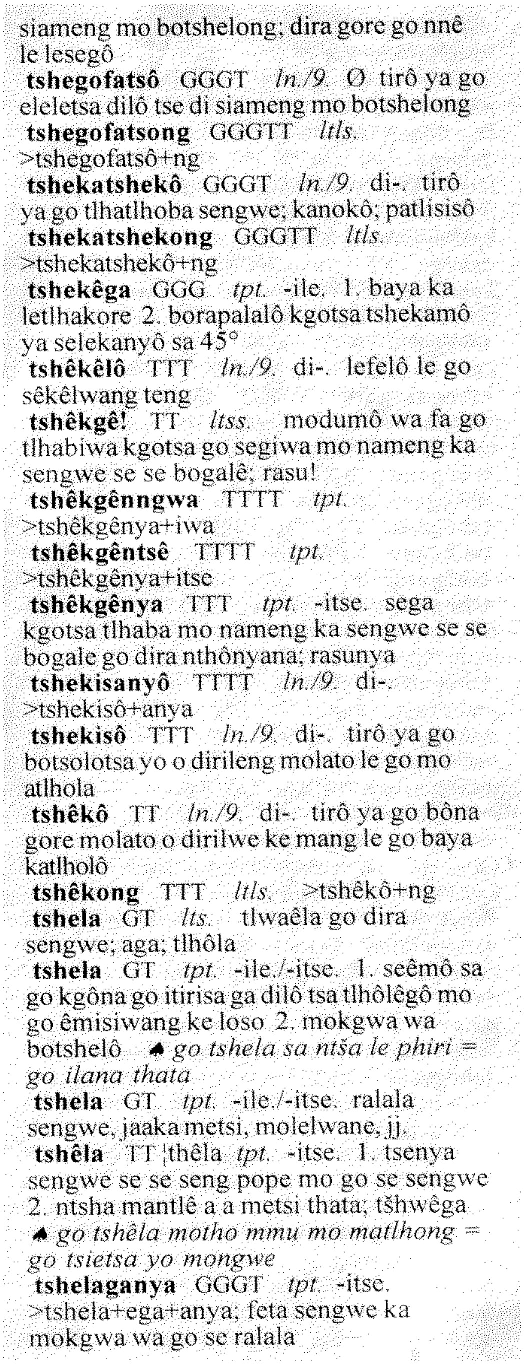
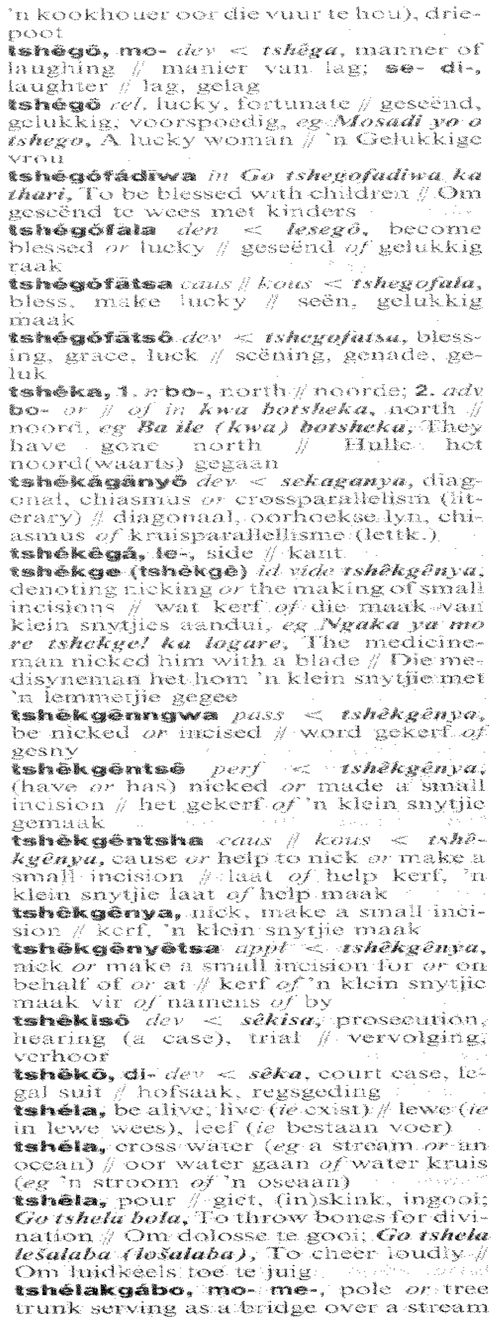
Although the THAN is directed at the native speakers of Setswana who have a strong command and knowledge of their language, users are also expected to have sufficient knowledge of morphology when it comes to nouns suffixes. This is often not the case. It is disappointing that noun inflections such as diminutives are omitted. Once again the question can be asked, on which grounds the items were selected.

4.6 Inconsistency regarding the lemmatization of homonyms

Homonyms are described as two or more unrelated meanings associated with the same form, for example, *bank* (of a river) and *bank* (financial entity). The treatment of homonyms in Setswana dictionaries also reveals some shortcomings. The question pertaining to the selection of entries is when to consider two occurrences of a word as having related or unrelated meanings. Compare the following examples 21 and 22 from (THAN and SEAD) and (THAN and THAND) respectively.

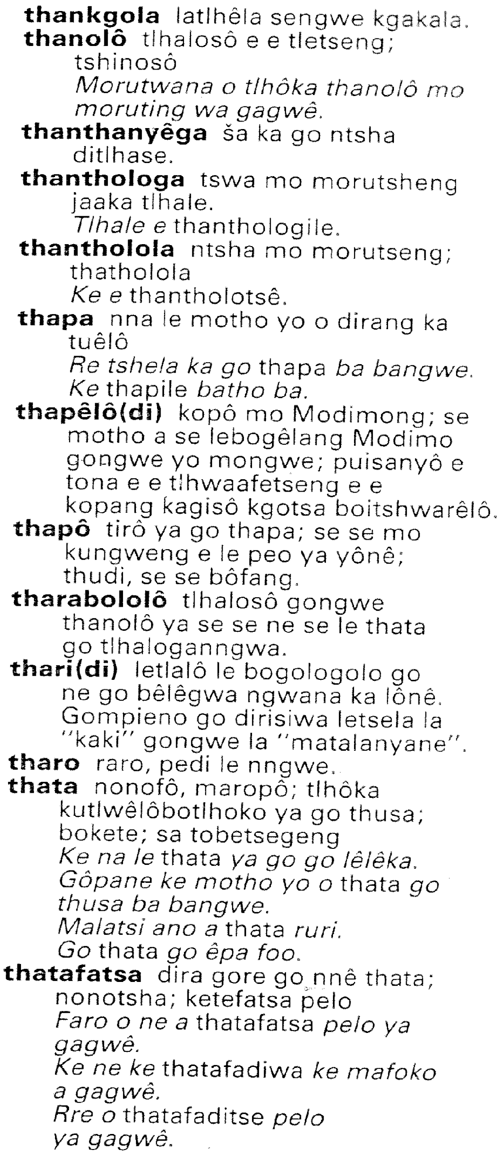
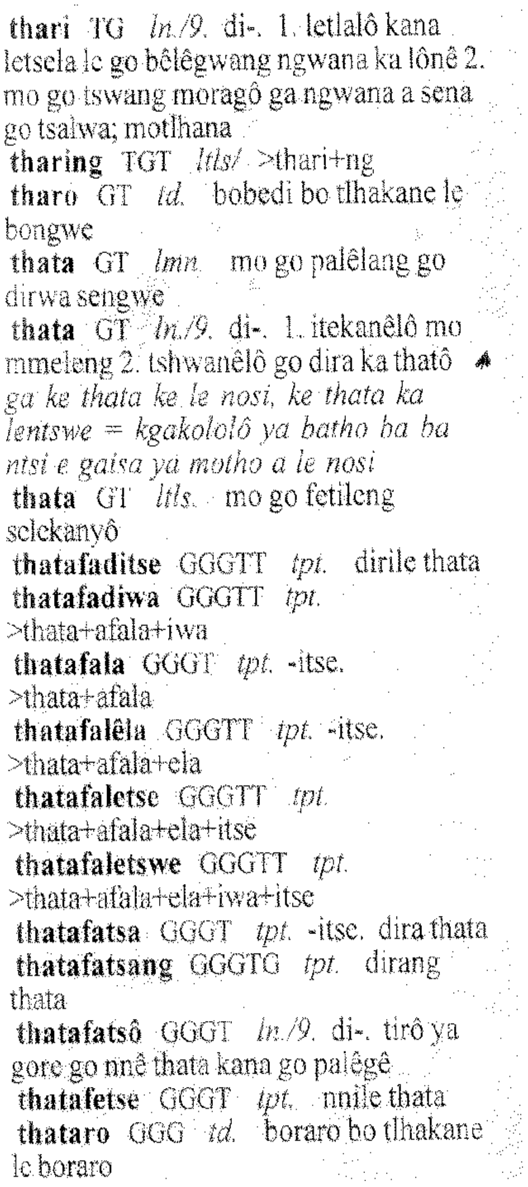
Example 21: tshela

THAN	SEAD


siameng mo botshelong; dira gore go nnê le lesegô
tshegofatsô GGGT *ln./9.* Ø tirô ya go eleletsa dilô tse di siameng mo botshelong
tshegofatsong GGGTT *ltls.* >tshegofatsô+ng
tshekatshekô GGGT *ln./9.* di-. tirô ya go tlhatlhoba sengwe; kanokô; patlisisô
tshekatshekong GGGTT *ltls.* >tshekatshekô+ng
tshekêga GGG *tpt.* -ile. 1. baya ka letlhakore 2. borapalalô kgotsa tshekamô ya selekanyô sa 45°
tshêkêlô TTT *ln./9.* di-. lefelô le go sêkêlwang teng
tshêkgê! TT *ltss.* modumô wa fa go tlhabiwa kgotsa go segiwa mo nameng ka sengwe se se bogalê; rasu!
tshêkgênngwa TTTT *tpt.* >tshêkgênya+iwa
tshêkgêntsê TTTT *tpt.* >tshêkgênya+itse
tshêkgênya TTT *tpt.* -itse. sega kgotsa tlhaba mo nameng ka sengwe se se bogale go dira nthônyana; rasunya
tshekisanyô TTTT *ln./9.* di-. >tshekisô+anya
tshekisô TTT *ln./9.* di-. tirô ya go botsolotsa yo o dirileng molato le go mo atlhola
tshêkô TT *ln./9.* di-. tirô ya go bôna gore molato o dirilwe ke mang le go baya katlholô
tshêkong TTT *ltls.* >tshêkô+ng
tshela GT *lts.* tlwaêla go dira sengwe; aga; tlhôla
tshela GT *tpt.* -ile./-itse. 1. seêmô sa go kgôna go itirisa ga dilô tsa tlhôlêgô mo go êmisiwang ke loso 2. mokgwa wa botshelô ♠ *go tshela sa ntša le phiri = go ilana thata*
tshela GT *tpt.* -ile./-itse. ralala sengwe, jaaka metsi, molelwane, jj.
tshêla TT ¦thêla *tpt.* -itse. 1. tsenya sengwe se se seng pope mo go se sengwe 2. ntsha mantlê a a metsi thata; tšhwêga ♠ *go tshêla motho mmu mo matlhong = go tsietsa yo mongwe*
tshelaganya GGGT *tpt.* -itse. >tshela+ega+anya; feta sengwe ka mokgwa wa go se ralala
	
	'n kookhouer oor die vuur te hou), driepoot
	tshêgô, mo- *dev* < *tshêga*, manner of laughing // manier van lag; **se- di-**, laughter // lag, gelag
	tshégô *rel*, lucky, fortunate // geseënd, gelukkig, voorspoedig, *eg Mosadi yo o tshego*, A lucky woman // 'n Gelukkige vrou
	tshégôfàdiwa *in Go tshegofadiwa ka thari*, To be blessed with children // Om geseënd te wees met kinders
	tshégôfàla *den* < *lesegô*, become blessed *or* lucky // geseënd *of* gelukkig raak
	tshégôfàtsa *caus* // *kous* < *tshegofala*, bless, make lucky // seën, gelukkig maak
	tshégôfàtsô *dev* < *tshegofatsa*, blessing, grace, luck // seëning, genade, geluk
	tshêka, 1. *n* **bo-**, north // noorde; **2.** *adv* **bo-** *or* // *of in kwa botsheka*, north // noord, *eg Ba ile (kwa) botsheka*, They have gone north // Hulle het noord(waarts) gegaan
	tshêkàgànyô *dev* < *sekaganya*, diagonal, chiasmus *or* crossparallelism (literary) // diagonaal, oorhoekse lyn, chiasmus *of* kruisparallellisme (lettk.)
	tshékêgá, le-, side // kant
	tshêkge (tshêkgê) *id vide tshêkgênya*, denoting nicking *or* the making of small incisions // wat kerf *of* die maak van klein snytjies aandui, *eg Ngaka ya mo re tshekge! ka logare*, The medicineman nicked him with a blade // Die medisyneman het hom 'n klein snytjie met 'n lemmetjie gegee
	tshêkgênngwa *pass* < *tshêkgênya*, be nicked *or* incised // word gekerf *of* gesny
	tshêkgêntsê *perf* < *tshêkgênya*, (have *or* has) nicked *or* made a small incision // het gekerf *of* 'n klein snytjie gemaak
	tshêkgêntsha *caus* // *kous* < *tshêkgênya*, cause *or* help to nick *or* make a small incision // laat *of* help kerf, 'n klein snytjie laat *of* help maak
	tshêkgênya, nick, make a small incision // kerf, 'n klein snytjie maak
	tshêkgênyêtsa *appl* < *tshêkgênya*, nick *or* make a small incision for *or* on behalf of *or* at // kerf *of* 'n klein snytjie maak vir *of* namens *of* by
	tshêkisô *dev* < *sêkisa*, prosecution, hearing (a case), trial // vervolging, verhoor
	tshêkô, di- *dev* < *sêka*, court case, legal suit // hofsaak, regsgeding
	tshéla, be alive, live (*ie* exist) // lewe (*ie* in lewe wees), leef (*ie* bestaan voer)
	tshéla, cross water (*eg* a stream *or* an ocean) // oor water gaan *of* water kruis (*eg* 'n stroom *of* 'n oseaan)
	tshêla, pour // giet, (in)skink, ingooi; *Go tshela bola*, To throw bones for divination // Om dolosse te gooi; *Go tshela lešalaba (lošalaba)*, To cheer loudly // Om luidkeels toe te juig
	tshélakgábo, mo- me-, pole *or* tree trunk serving as a bridge over a stream

From example 21 above, THAN provides treatment of homonyms *tshela* as *tshela* (used to), *tshêla* (pour), *tshela* (cross) and *tshela* (to live) while SEAD provides only three homonyms i.e. *tshela* (to live), *tshela* (to cross) and *tshêla* (pour). Consider also the second example of *thari* taken from the THAN and the THAND:

Example 22: thari

THAN	THAND
 **thankgola** latlhêla sengwe kgakala. **thanolô** tlhalosô e e tletseng; tshinosô *Morutwana o tlhôka thanolô mo moruting wa gagwê.* **thanthanyêga** ša ka go ntsha ditlhase. **thanthologa** tswa mo morutsheng jaaka tlhale. *Tlhale e* thanthologile. **thantholola** ntsha mo morutseng; thatholola *Ke e* thantholotsê. **thapa** nna le motho yo o dirang ka tuêlô *Re tshela ka go* thapa *ba bangwe.* *Ke* thapile *batho ba.* **thapêlô(di)** kopô mo Modimong; se motho a se lebogêlang Modimo gongwe yo mongwe; puisanyô e tona e e tlhwaafetseng e e kopang kagisô kgotsa boitshwarêlô. **thapô** tirô ya go thapa; se se mo kungweng e le peo ya yônê; thudi, se se bôfang. **tharabololô** tlhalosô gongwe thanolô ya se se ne se le thata go tlhaloganngwa. **thari(di)** letlalô le bogologolo go ne go bêlêgwa ngwana ka lônê. Gompieno go dirisiwa letsela la "kaki" gongwe la "matalanyane". **tharo** raro, pedi le nngwe. **thata** nonofô, maropô; tlhôka kutlwêlôbotlhoko ya go thusa; bokete; sa tobetsegeng *Ke na le* thata *ya go go lêlêka.* *Gôpane ke motho yo o* thata *go thusa ba bangwe.* *Malatsi ano a* thata *ruri.* *Go* thata *go êpa foo.* **thatafatsa** dira gore go nnê thata; nonotsha; ketefatsa pelo *Faro o ne a* thatafatsa *pelo ya gagwê.* *Ke ne ke* thatafadiwa *ke mafoko a gagwê.* *Rre o* thatafaditse *pelo ya gagwê.*	 **thari** TG *ln./9.* di-. 1. letlalô kana letsela le go bêlêgwang ngwana ka lônê 2. mo go tswang moragô ga ngwana a sena go tsalwa; motlhana **tharing** TGT *ltls/* >thari+ng **tharo** GT *td.* bobedi bo tlhakane le bongwe **thata** GT *lmn.* mo go palêlang go dirwa sengwe **thata** GT *ln./9.* di-. 1. itekanêlô mo mmeleng 2. tshwanêlô go dira ka thatô *ga ke thata ke le nosi, ke thata ka lentswe = kgakololô ya batho ba ba ntsi e gaisa ya motho a le nosi* **thata** GT *ltls.* mo go fetileng selekanyô **thatafaditse** GGGTT *tpt.* dirile thata **thatafadiwa** GGGTT *tpt.* >thata+afala+iwa **thatafala** GGGT *tpt.* -itse. >thata+afala **thatafalêla** GGGTT *tpt.* -itse. >thata+afala+ela **thatafaletse** GGGTT *tpt.* >thata+afala+ela+itse **thatafaletswe** GGGTT *tpt.* >thata+afala+ela+iwa+itse **thatafatsa** GGGT *tpt.* -itse. dira thata **thatafatsang** GGGTG *tpt.* dirang thata **thatafatsô** GGGT *ln./9.* di-. tirô ya gore go nnê thata kana go palêgê **thatafetse** GGGT *tpt.* nnile thata **thataro** GGG *td.* boraro bo tlhakane le boraro

From example 22 above, both the THAN and the THAND managed to provide only one sense of the homonym *thari* as (skin used to carry a baby) while other senses are excluded.

From examples 21 and 22 for *tshela* and *thari*, it is clear that frequency counts and user-friendliness were not taken into consideration. These are words which are frequently used. The problem with frequency counts by corpus query programs such as WordSmith Tools is that a single total count is given for the orthographic form shared by both homonyms. The lexicographer has to determine the frequencies manually. It is important to emphasize that the distinction between the homonyms be determined on the basis of frequency counts and that the structural markers be employed to indicate the different contexts in which the lemma signs *tshela* and *thari* can occur. The above mentioned statement is supported by Hausman and Wiegand (1989:356) where they emphasize the use of the structural markers to demonstrate the different meanings of homonyms and maintain that these should not be presented haphazardly but according to a predetermined set of criteria accounted for in the front matter text. Consider the following example in this regard:

Example 23

Tshêla^{1}	(to put in)
Tshéla^{2}	(to cross)
Tshéla^{3}	(to live)

Compare also the following examples where Afrikaans and English dictionaries are consistently managing to lemmatize the homonyms:

Example 24: Verklarende Afrikaanse WoordeBoek (VAW)

Graaf^1, (s), **grawe**. **1**. Adellike titel. **2.** Iem. Met so 'n titel.

Graaf^2, (s), **grawe. 1**. Spitwerking. **2**. Lengte van 'n graafblad. **3**. Hoeveelheid wat op 'n keer met 'n graaf geskep kan word; ~ steel.

Example 25: TSAOSD

lag^1 *verb* (**lagged, lagging**) go too slowly and fail to keep up with others.

lag^2 *noun* lagging; a delay.

lag^3 *verb* (**lagged, lagging**) wrap pipes or boilers etc. in insulating material to keep them warm.

4.7 The absence of tonal indication

Tonal indication is of crucial importance in Setswana for it is used to make both semantic and grammatical distinction. As far as tonal indication is concerned, it is important to note that dictionaries employ different conventions. Some dictionaries indicate high as well as low tones while others indicate only high tones. For example, Ziervogel enters head words or sub-head words in capital letters with indication of high tone e.g. *RÉKA*. The word is repeated in brackets with the circumflexes indicated on *e* or *o*: *rêka*. For such entries he uses lower case: *lerêko*, *ma- rêko*. A second option is to indicate the tonal pattern separately by means of the upper case character **L** for low tone and **H** for high tone. For example, Van Wyk in Pukuntšu enters head words with the circumflexes indicated, as follows: *boikêtsiso* and indicates the tonal pattern further down in the article as, for example, LLHHL.

As far as the Setswana dictionaries are concerned, it is unfortunate to state that only SEAD succeeded in showing tonal indication. In this dictionary, a distinction is made between a high, a low and a falling tone which are indicated by means of two tonal diacritics i.e. (′) high tone, (¯) a falling tone while the low tone is left unmarked. In the dictionary the mid-low vowels [ε] and [Φ] are marked with a circumflex, for example *ê* in *êma* and *ô* in *ôma*. One gets the impression that compilers of the SEAD regarded tonal information of equal importance to lexical information. It is disappoinding that other Setswana dictionaries such as THAN, THAND, MSED and SED excluded tonal distinction. The example below will now be used to demonstrate the importance of tonal indication in Setswana dictionaries.

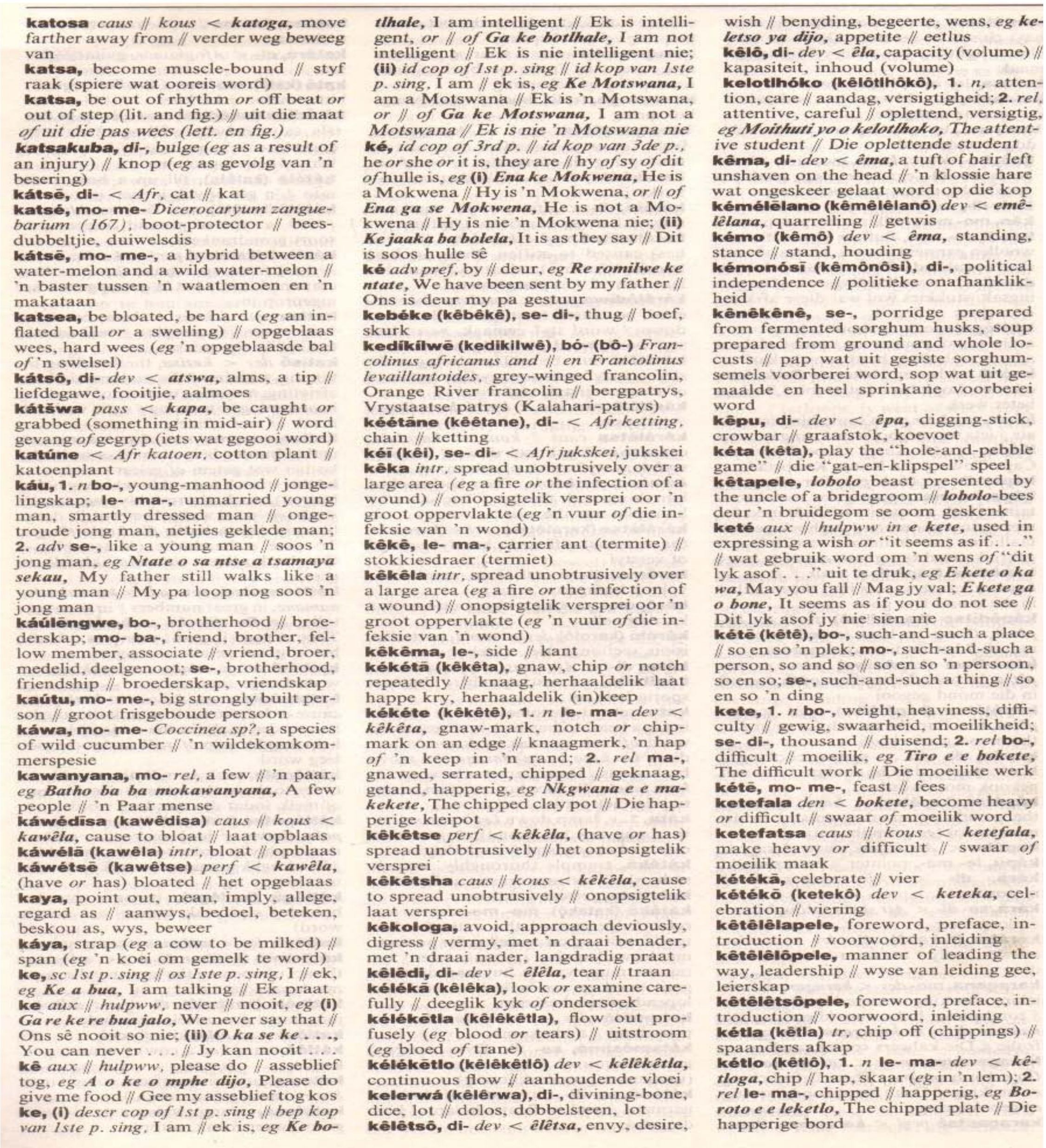
Example 26

Item **240** in the Setswana lemmatized frequency list is ***tshela***. Without tonal indication, this form could mean any of the three possibilities listed below:

1.	*tshéla*^{1}	[verb]	‘to jump’
2.	*tshêla*^{2}	[verb]	‘to pour’
3.	*tshéla*^{3}	[verb]	‘to cross’

It is suggested that tonal indication be included in the forthcoming Setswana dictionaries to render a more user-friendly Setswana dictionary. Compare now an example extracted from SEAD where compilers consistently strive to make use of tonal indication.

Example 27: SEAD



katosa *caus* // *kous* < ***katoga,*** move farther away from // verder weg beweeg van

katsa, become muscle-bound // styf raak (spiere wat ooreis word)

katsa, be out of rhythm *or* off beat *or* out of step (lit. and fig.) // uit die maat *of uit die pas wees (lett. en fig.)*

katsakuba, di-, bulge (*eg* as a result of an injury) // knop (*eg* as gevolg van 'n besering)

kátsē, di- < *Afr*, cat // kat

katsé, mo- me- *Dicerocaryum zanguebarium (167)*, boot-protector // beesdubbeltjie, duiwelsdis

kátsē, mo- me-, a hybrid between a water-melon and a wild water-melon // 'n baster tussen 'n waatlemoen en 'n makataan

katsea, be bloated, be hard (*eg* an inflated ball *or* a swelling) // opgeblaas wees, hard wees (*eg* 'n opgeblaasde bal *of* 'n swelsel)

kátsô, di- *dev* < ***atswa,*** alms, a tip // liefdegawe, fooitjie, aalmoes

kátšwa *pass* < ***kapa,*** be caught *or* grabbed (something in mid-air) // word gevang *of* gegryp (iets wat gegooi word)

katúne < *Afr katoen*, cotton plant // katoenplant

káu, 1. *n* **bo-,** young-manhood // jongelingskap; **le- ma-,** unmarried young man, smartly dressed man // ongetroude jong man, netjies geklede man; **2.** *adv* **se-,** like a young man // soos 'n jong man, *eg* ***Ntate o sa ntse a tsamaya sekau,*** My father still walks like a young man // My pa loop nog soos 'n jong man

káúlēngwe, bo-, brotherhood // broederskap; **mo- ba-,** friend, brother, fellow member, associate // vriend, broer, medelid, deelgenoot; **se-,** brotherhood, friendship // broederskap, vriendskap

kaútu, mo- me-, big strongly built person // groot frisgeboude persoon

káwa, mo- me- *Coccinea sp?*, a species of wild cucumber // 'n wildekomkommerspesie

kawanyana, mo- *rel*, a few // 'n paar, *eg* ***Batho ba ba mokawanyana,*** A few people // 'n Paar mense

káwédīsa (kawêdisa) *caus* // *kous* < ***kawêla***, cause to bloat // laat opblaas

káwélā (kawêla) *intr*, bloat // opblaas

káwétsē (kawêtse) *perf* < ***kawêla,*** (have *or* has) bloated // het opgeblaas

kaya, point out, mean, imply, allege, regard as // aanwys, bedoel, beteken, beskou as, wys, beweer

káya, strap (*eg* a cow to be milked) // span (*eg* 'n koei om gemelk te word)

ke, *sc 1st p. sing* // *os 1ste p. sing*, I // ek, *eg* ***Ke a bua,*** I am talking // Ek praat

ke *aux* // *hulpww*, never // nooit, *eg* **(i)** ***Ga re ke re bua jalo,*** We never say that // Ons sê nooit so nie; **(ii)** ***O ka se ke . . .,*** You can never . . . // Jy kan nooit . . .

kê *aux* // *hulpww*, please do // asseblief tog, *eg* ***A o ke o mphe dijo,*** Please do give me food // Gee my asseblief tog kos

ke, (i) *descr cop of 1st p. sing* // *bep kop van 1ste p. sing*, I am // ek is, *eg* ***Ke botlhale,*** I am intelligent // Ek is intelligent, *or* // *of* ***Ga ke botlhale,*** I am not intelligent // Ek is nie intelligent nie; **(ii)** *id cop of 1st p. sing* // *id kop van 1ste p. sing*, I am // ek is, *eg* ***Ke Motswana,*** I am a Motswana // Ek is 'n Motswana, *or* // *of* ***Ga ke Motswana,*** I am not a *Motswana // Ek is nie 'n Motswana nie*

ké, *id cop of 3rd p.* // *id kop van 3de p.*, he *or* she *or* it is, they are // hy *of* sy *of* dit *of* hulle is, *eg* **(i)** ***Ena ke Mokwena,*** He is a Mokwena // Hy is 'n Mokwena, *or* // *of* ***Ena ga se Mokwena,*** He is not a Mokwena // Hy is nie 'n Mokwena nie; **(ii)** ***Ke jaaka ba bolela,*** It is as they say // Dit is soos hulle sê

ké *adv pref*, by // deur, *eg* ***Re romilwe ke ntate,*** We have been sent by my father // Ons is deur my pa gestuur

kebéke (kêbêkê), se- di-, thug // boef, skurk

kedíkílwē (kedikilwê), bó- (bô-) *Francolinus africanus and // en Francolinus levaillantoides*, grey-winged francolin, Orange River francolin // bergpatrys, Vrystaatse patrys (Kalahari-patrys)

kéétāne (kêêtane), di- < *Afr ketting*, chain // ketting

kéī (kêi), se- di- < *Afr jukskei*, jukskei

kêka *intr*, spread unobtrusively over a large area (*eg* a fire *or* the infection of a wound) // onopsigtelik versprei oor 'n groot oppervlakte (*eg* 'n vuur *of* die infeksie van 'n wond)

kêkê, le- ma-, carrier ant (termite) // stokkiesdraer (termiet)

kêkêla *intr*, spread unobtrusively over a large area (*eg* a fire *or* the infection of a wound) // onopsigtelik versprei oor 'n groot oppervlakte (*eg* 'n vuur *of* die infeksie van 'n wond)

kêkêma, le-, side // kant

kékétā (kêkêta), gnaw, chip *or* notch repeatedly // knaag, herhaaldelik laat happe kry, herhaaldelik (in)keep

kékéte (kêkêtê), 1. *n* **le- ma-** *dev* < ***kêkêta,*** gnaw-mark, notch *or* chipmark on an edge // knaagmerk, 'n hap *of* 'n keep in 'n rand; **2.** *rel* **ma-,** gnawed, serrated, chipped // geknaag, getand, happerig, *eg* ***Nkgwana e e makekete,*** The chipped clay pot // Die happerige kleipot

kêkêtse *perf* < ***kêkêla,*** (have *or* has) spread unobtrusively // het onopsigtelik versprei

kêkêtsha *caus* // *kous* < ***kêkêla,*** cause to spread unobtrusively // onopsigtelik laat versprei

kêkologa, avoid, approach deviously, digress // vermy, met 'n draai benader, met 'n draai nader, langdradig praat

kêlêdi, di- *dev* < ***êlêla,*** tear // traan

kéléká (kêlêka), look *or* examine carefully // deeglik kyk *of* ondersoek

kéléketla (kêlêkêtla), flow out profusely (*eg* blood *or* tears) // uitstroom (*eg* bloed *of* trane)

kéléketlo (kêlêkêtlô) *dev* < ***kêlêkêtla,*** continuous flow // aanhoudende vloei

kelerwá (kêlêrwa), di-, divining-bone, dice, lot // dolos, dobbelsteen, lot

kêlêtsô, di- *dev* < ***êlêtsa,*** envy, desire, wish // benyding, begeerte, wens, *eg* ***keletso ya dijo,*** appetite // eetlus

kêlô, di- *dev* < ***êla,*** capacity (volume) // kapasiteit, inhoud (volume)

kelotlhóko (kêlôtlhôkô), 1. *n*, attention, care // aandag, versigtigheid; **2.** *rel*, attentive, careful // oplettend, versigtig, *eg Moithuti yo o kelotlhoko, The attentive student // Die oplettende student*

kêma, di- *dev* < ***êma,*** a tuft of hair left unshaven on the head // 'n klossie hare wat ongeskeer gelaat word op die kop

kémélēlano (kêmêlêlanô) *dev* < ***emêlêlana,*** quarrelling // getwis

kémo (kêmô) *dev* < ***êma,*** standing, stance // stand, houding

kémonósī (kêmônôsi), di-, political independence // politieke onafhanklikheid

kênêkênê, se-, porridge prepared from fermented sorghum husks, soup prepared from ground and whole locusts // pap wat uit gegiste sorghumsemels voorberei word, sop wat uit gemaalde en heel sprinkane voorberei word

kêpu, di- *dev* < ***êpa,*** digging-stick, crowbar // graafstok, koevoet

kéta (kêta), play the "hole-and-pebble game" // die "gat-en-klipspel" speel

kêtapele, ***lobolo*** beast presented by the uncle of a bridegroom // ***lobolo***-bees deur 'n bruidegom se oom geskenk

keté *aux* // *hulpww in* ***e kete,*** used in expressing a wish *or* "it seems as if . . ." // wat gebruik word om 'n wens *of* "dit lyk asof . . ." uit te druk, *eg* ***E kete o ka wa,*** May you fall // Mag jy val; ***E kete ga o bone,*** It seems as if you do not see // Dit lyk asof jy nie sien nie

kétē (kêtê), bo-, such-and-such a place // so en so 'n plek; **mo-,** such-and-such a person, so and so // so en so 'n persoon, so en so; **se-,** such-and-such a thing // so en so 'n ding

kete, 1. *n* **bo-,** weight, heaviness, difficulty // gewig, swaarheid, moeilikheid; **se- di-,** thousand // duisend; **2.** *rel* **bo-,** difficult // moeilik, *eg* ***Tiro e e bokete,*** The difficult work // Die moeilike werk

kétē, mo- me-, feast // fees

ketefala *den* < ***bokete,*** become heavy *or* difficult // swaar *of* moeilik word

ketefatsa *caus* // *kous* < ***ketefala,*** make heavy *or* difficult // swaar *of* moeilik maak

kétékā, celebrate // vier

kétékō (ketekô) *dev* < ***keteka,*** celebration // viering

kêtêlêlapele, foreword, preface, introduction // voorwoord, inleiding

kêtêlêlôpele, manner of leading the way, leadership // wyse van leiding gee, leierskap

kêtêlêtsôpele, foreword, preface, introduction // voorwoord, inleiding

kétla (kêtla) *tr*, chip off (chippings) // spaanders afkap

kétlo (kêtlô), 1. *n* **le- ma-** *dev* < ***kêtloga,*** chip // hap, skaar (*eg* in 'n lem); **2.** *rel* **le- ma-,** chipped // happerig, *eg* ***Boroto e e leketlo,*** The chipped plate // Die happerige bord

In example 27 different meanings of *kete* are clearly distinguished on the basis of tonal indication. If such distinctions were not given the user would find it difficult to retrieve the correct information.

4.8 Dialect words as lemmas in the dictionary are evaluated against the existing Setswana dictionaries

The Setswana language has eight dialects. Dictionaries such as THAN, THAND, MSED and SED do not cover these regional varieties. They are only limited to one dialect i.e. the Serolong dialect which is considered as the standard language.

Wikipedia (2007) defines a dialect as a variety of a language used by people from a particular geographical area. Anthropological linguists on the other hand define a dialect as a specific form of a language used by a speech community. For example, in Setswana the sound represented by ***f*** *is* unknown to the Batlhaping tribe but very common among the Barolong tribe. The ***h*** in the word *le**h**atshe* (earth) is pronounced as **f** in Serolong as *le**f**atshe*. The glottal sound **g** of the southern Batswana tribe loses its glottal sound among the more northern tribes and becomes softened into the sound ***h*** and the sound ***sh*** becomes a pure ***s***. Compare the following examples in this regard:

Example 28

Alternates ***f***, ***g*** and ***h***

English	**Setlhaping**	**Serolong**	**Standard Setswana**
Earth	*le**h**atshe*	*le**f**atshe*	*le**f**atshe*
Cow	*k**h**omo*	*k**g**omo*	*k**g**omo*
Fat	*ma**h**ura*	*ma**h**ura*	*ma**f**ura*
Scares	*tlhoka**h**ala*	*tlhoka**g**ala*	*tlhoka**f**ala*
Rich	***f**uma/huma*	***g**uma*	***h**uma*
Found	***f**umane*	***f**itlhela*	*bone*

Compare other examples below where the vowel *o* in *lo-* is frequently replaced by *e* in *le-* with a possessive *la*.

Example 29

L**eriba* ***la *noka*	***L**oriba* ***lwa*** *noka*	(bank of the river)

Le*silo* ***la*** *mosimane*	***Lo****silo* ***lwa*** *mosimane*	(a stupid boy)
Le*sogodi* ***la*** *ditsie*	***Lo****sogodi* ***lwa*** *ditsie*	(a swarm of locusts)

Consider other examples taken from a Botswana Television conversation:

Example 30

English	**Botswana T.V.**	**Sekgatla**	**Sekwena**	**Standard**
beat	***betsa***	*itaya/šapa*	*otla*	*betsa*
school	***sekwele***	*sekolo*	*sekolo*	*sekolo*
week	***biki***	*beke*	*beke*	*beke*
huge	***setona***	*setona/segolo*	*segolo*	*segolo*

If one compares the Botswana Television conversation with the standard language, one realises that we do not speak of dialects but alternates patterns since the dialectical boundries are broken up by the population shifts, urbanized, political re-organisation and technology.

Consider another example of alternates below:

Example 31

Alternates tš, ts, tšh and tšhwa

English	**Serolong**	**Sekgatla**	**Sekwena**	**Standard**
dog	*ntsha*	*mpša*	*ntšwa/ntšha*	*ntšhwa*
ostrich	*mpshe*	*mpše*	*ntšhe*	*ntšhwe*
youths	*batšha*	*bašwa*	*baša*	*baswa*

The lexicographer will have to justify the alternants in deciding whether to include or exclude in the macrostructure of the dictionary. Consider cojuctions as other examples of alternants:

Example 32

Conjuctions

English	**Serolong**	**Sengwaketse**	**Setlhaping**	**Standard**
though	***f****ela*	***g****ela*	***h****ela*	***f****ela*
if	***f****a*	***g****e*	***h****a*	***f****a*

It is important to note that (*f*) is pronounced as (*g*) in Sengwaketse and as (*h*) in Setlhaping. Consider the following example sentences:

> *Ke tla bua* ***fela fa*** *o ka ntetla* (I will speak only if you allow me)
> *Ke tla bua* ***gela ge*** *o ka ntetla* (I will speak only if you allow me)
> *Ke tla bua* ***hela ha*** *o ka ntetla.* (I will speak only if you allow me)

According to the research, the ***f*** in Setswana is favoured more than the ***h***. Singled it can be justifiable. The ***h*** is a matter of Southern Sotho influence which is also affected by the geographical regions. Consider other examples taken from the Sekwena dialect:

Example 33

English	**Sekwena**	**Standard**
I don’t know	*kgitse*	*ga ke itse*
I have been	*kgebolo*	*ga ke bolo*
I don’t want	*kgebatle*	*ga ke battle*
Here	*kweno*	*kwano*

If one compares the Sekwena dialects with the standard language from example 33, it is important to note that the negative morpheme *ga ke* (I don't) has been coined to the verb stem itse, *bolo* and *battle*. Consder the following:

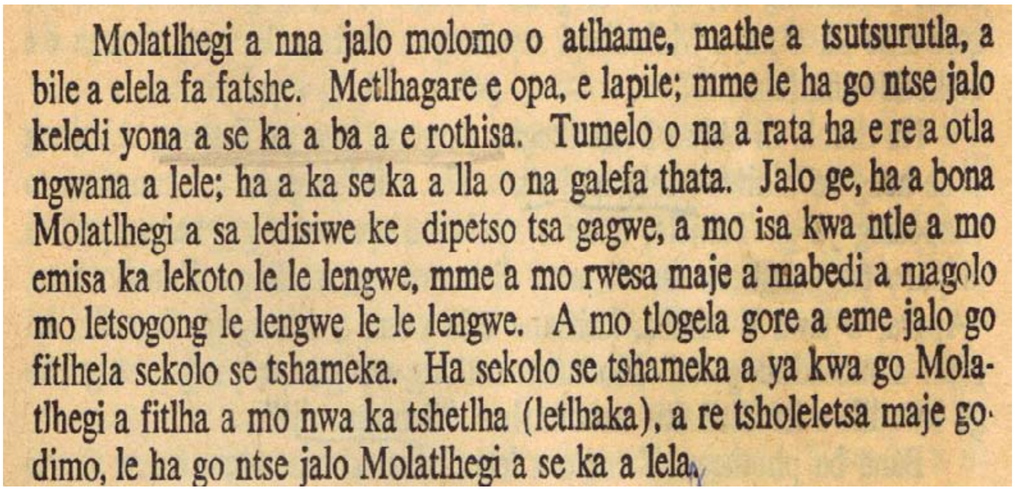
Kg + itse	*> kgitse*
Kg + bolo	*> kgebolo*
Kg + battle	*> kgebatle*

Consider other examples taken from selected forms of the Setswana drama books and novels:

Example 34

Selected passages from dramas and novels

Motimedi: D.P. Moloto



Molatlhegi a nna jalo molomo o atlhame, mathe a tsutsurutla, a bile a elela fa fatshe. Metlhagare e opa, e lapile; mme le ha go ntse jalo keledi yona a se ka a ba a e rothisa. Tumelo o na a rata ha e re a otla ngwana a lele; ha a ka se ka a lla o na galefa thata. Jalo ge, ha a bona Molatlhegi a sa ledisiwe ke dipetso tsa gagwe, a mo isa kwa ntle a mo emisa ka lekoto le le lengwe, mme a mo rwesa maje a mabedi a magolo mo letsogong le lengwe le le lengwe. A mo tlogela gore a eme jalo go fitlhela sekolo se tshameka. Ha sekolo se tshameka a ya kwa go Molatlhegi a fitlha a mo nwa ka tshetlha (letlhaka), a re tsholeletsa maje godimo, le ha go ntse jalo Molatlhegi a se ka a lela.

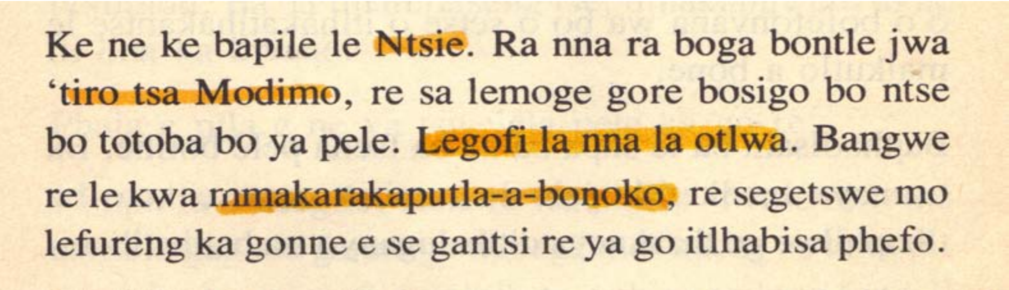
From the above example, it is clear that the author used different dialects and languages as alternates. For example, the use of '*ha*' instead of '*ga*' which is the direct influence from the Setlhaping dialect; the use of '*otla*' (beat) instead of *betsa* or *itaya* which is the direct influence from the Sekwena dialect and the use of the word *lekoto* (leg) which derives from the Sepedi language.

Moratho o montsho: S.S. Tshetlho

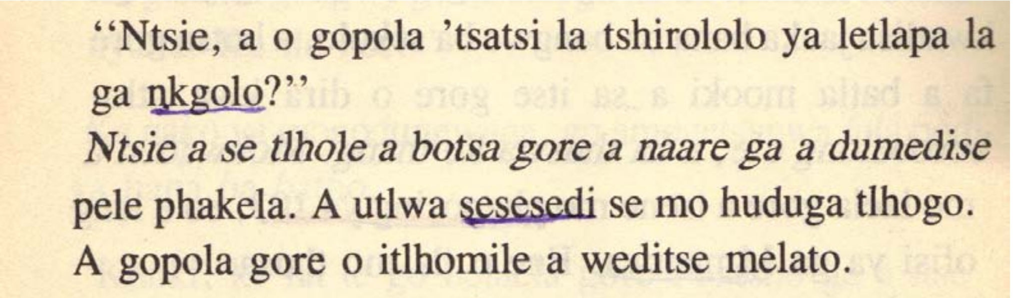
Ka ikutlwa ke otlwa ke letswalo. Ke tsamaiwa ke phefo e e maruru mo mokwatleng ke le mosadi. Ya re a laela, modumo wa tlhatloga le go feta.

From the above given example, the author makes use of the word '*otlwa*' (beaten) instead of *betswa* or *itawa* which is the direct influence from the Sekwena dialect.

Consider another example below:



Ke ne ke bapile le Ntsie. Ra nna ra boga bontle jwa 'tiro tsa Modimo, re sa lemoge gore bosigo bo ntse bo totoba bo ya pele. Legofi la nna la otlwa. Bangwe re le kwa mmakarakaputla-a-bonoko, re segetswe mo lefureng ka gonne e se gantsi re ya go itlhabisa phefo.



"Ntsie, a o gopola 'tsatsi la tshirololo ya letlapa la ga nkgolo?"

Ntsie a se tlhole a botsa gore a naare ga a dumedise pele phakela. A utlwa sesesedi se mo huduga tlhogo. A gopola gore o itlhomile a weditse melato.

In this example the author makes use of the word '*nkgolo*' (grandfather) instead of *rremogolo* which derives from the Southern Sotho language and uses the sound '*h*' instead of '*g*' which is the direct influence from the Serolong dialect.

From the above given examples, one notices the inconsistencies of the contemporary author's orthography i.e. authors do not spell the words in the same way which indicate that there is no standard writing in the field.

It is important to note that lexicographers should look at a dialect as a bundle of characteristics peculiar to a language in a specific social environment. The lexicographer should keep in mind that a dictionary should be fully descriptive. It should record objectively various dialects and different styles. However this is not the case with the Setswana dictionaries. It is important to state that future Setswana dictionaries should pay more attention to dialectical forms. Although the whole issue regarding Setswana dialects is complicated, at least frequently used dialectical forms should be given and appropriately labelled.

4.9 Conclusion

In this chapter the extent to which the respective Setswana dictionaries succeeded or failed to treat the most commonly used words was indicated. We have critically analysed and evaluated the typical macrostructural inconsistencies existing in Setswana dictionaries. Macrostructural aspects relating to the inclusion versus the omission of individual lemmata was dealt with. In addition it was also stressed how corpora can be put to good use in revising and improving the macrostructure of the existing Setswana dictionaries. The unequal treatment of derived forms of verbs which results from a lemmatisation approach where lexicographers simply add words as they come across them was also illustrated. Key components of the revision strategy including the design and the use of a multi-dimensional Ruler and Block System for the measurement and balancing of the alphabetical stretches in terms of number of pages per alphabetical category was also highlighted. In addition, the importance of the relationship existing between frequencies and dictionaries was also emphasized. It is also important to state why the focus is on the high frequencies and not on the low frequencies. For example, for bilinguals, given the restriction on the number of pages top frequencies can hardly be covered and that our tests have proven that the top 10, 000 frequencies cover more than 95% of Setswana texts. As for monolingual we try to include lower frequencies of special relevance such as cultural terms but given the users from small children to adults we cannot dare to omit the top frequencies. In this chapter, we have also provided a perspective on how the South African Bantu language lexicography reflects a complex interplay of lemmatisation traditions, lemmatisation strategies and lemmatisation approaches.

The importance of tonal indication in Setswana where tones are used to make both semantic and grammatical distinctions was illustrated by means of suitable examples taken from the two monolingual Setswana dictionaries i.e. THAN and THAND. Problems regarding the lemmatization of the noun prefixes *le-* and *lo-*, the absence of nominal derivations and the absence of the treatment of homonyms in the Setswana dictionaries were highlighted and critically analysed. The evaluation of dialect words as lemmas in Setswana dictionaries was demonstrated and critically analysed to determine whether the Setswana is standardised or not.

It is also important to note that the dictionary situation in Setswana is such that we do not have the luxury of compiling dictionaries for narrowly defined target users, separate dictionaries for productive and receptive use or dictionaries for the different dialects. To date the Setswana lexicographer was forced to compile general dictionaries for use by everyone and to include the most relevant dialectical forms. On the question, do Setswana dictionaries need a descriptive dictionary or a normative one, one could say in principle descriptive but also normative aspects since the language is not fully standardized.

Chapter 5

The microstructure

5.1 Introduction

This chapter gives an explanation of how corpora can be seen as a key to writing better Setswana dictionary articles on the microstructural level. In this chapter, we will discuss three major issues. Firstly, we will cover the importance of corpora in the following areas; as sense distinctions, as a key to writing better dictionary articles, as an aid to retrieve typical collocations and corpora as an aid to select typical and natural examples. Secondly, we will highlight certain microstructural inconsistencies relating to the treatment of verbs and 'the so-called' Setswana synonyms. Lastly, the treatment of the Setswana months in the currently available Setswana dictionaries will be critically analysed and evaluated against the background information of the English and the Afrikaans dictionaries. Each section will conclude with suggestions for the improvement of the respective Setswana dictionaries by means of a corpus-based microstructure.

It is argued that, if African-language lexicography is to take its rightful place in the new millennium, the active use of corpora to improve the quality of microstructural elements in the treatment of lemma signs should become an absolute priority. (cf. De Schryver and Prinsloo (2000a and 2000b) and Prinsloo and De Schryver (1999 and 2001). Corpora provide useful evidence of the formal usage of the lexical items, i.e. the associated syntactic structures, pharaseological patterns, collocations, contexts of use, etc.

According to Galley (2000:132), the microstructure should include a diverse mass of data, for example; cross-references, paraphrase of meaning, examples, parts of speech, typographical exposition, to mention but a few. The basic aim of the lexicographer is to guide the user in respect of the properties/features, characteristics, use and meaning of a lemma sign. Laufer, (1992:71) formulates this basic aim as follows:

> "Knowing a word would ideally imply familiarity with all its propertie […] When a person 'knows' a word, he/she knows the following: the word's pronunciation, its spelling, its morphological components if any, the words that are morphologically related to it, the word's syntactic behaviour in a sentence, the full range of the word's meaning, the appropriate situations for using the word, it's collocation restrictions, its distribution and the relation between the word and other words within a lexical set".

The question that arises now is how the utilization of a corpus can help the lexicographer to achieve the ultimate microstructural goal. According to De Schryver and Prinsloo (2000), a large, structured, electronic corpus is the first requirement for corpus-based dictionaries as well as advanced corpus query tools. Such tools must be able to provide at least two basic outputs, namely word-frequency counts and concordance lines as well as the capacity of analysing problematic contexts. Concordance lines culled from living-language sources supplement and support the lexicographer's (native-speaker) intuition. They take him/her to the heart of the actual usage of word(s) in context, allowing the lexicographer to see up to several dozens of contexts at a glance.

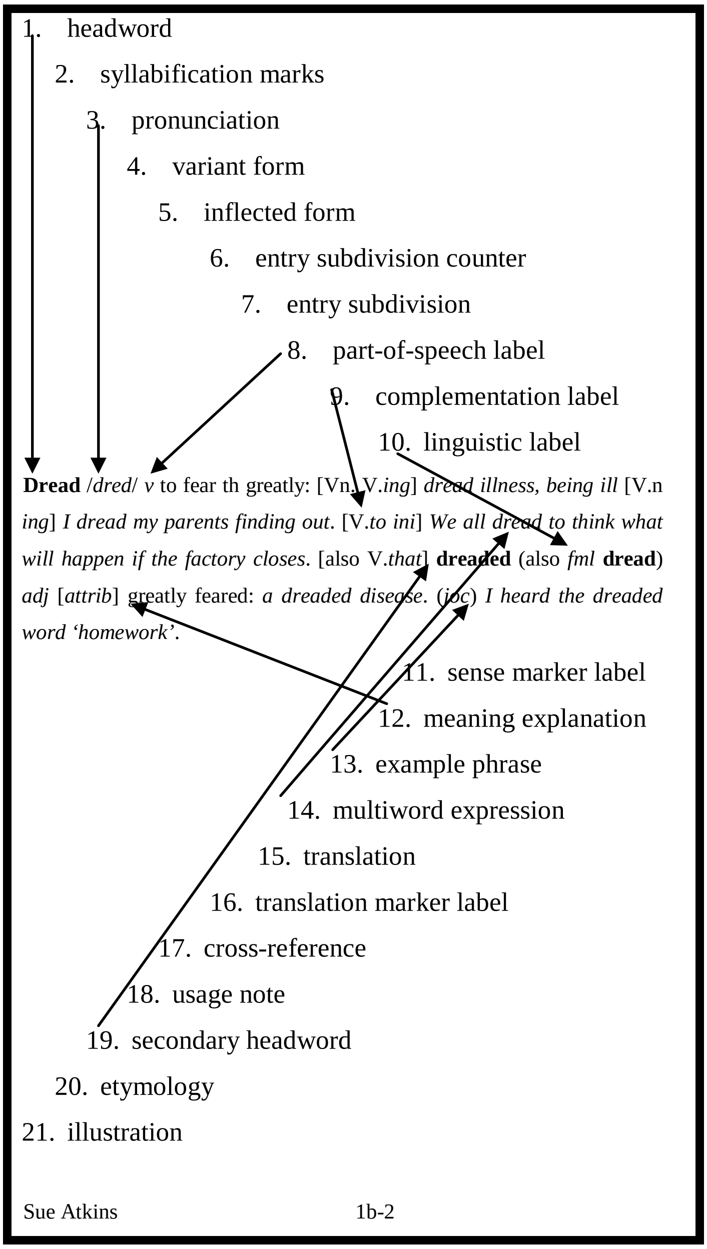
In order to illustrate this interaction between corpus queries and the compilation of a dictionary's microstructure, the chapter will be structured as follows: First, a brief introduction to corpus queries as an aid to sense distinctions is given with reference to the Setswana homonyms such as, *tshela* and *thari*. This is followed by detailed analysis of corpus lines in combination with frequency counts for the so-called synonyms such as *batla* (to look for) and *senka* (to look for). Finally, inconsistencies of application in the treatment of verbs and the Setswana months within the microstructure of the currently available Setswana dictionaries will be highlighted.

5.2 Inconsistencies of application within the microstructure of the Setswana dictionaries

On the microstructural level, comment on semantics is the most important data type. Gouws (1983:113) states that it is the information type most generally consulted by the target users, most substantial and considered as the central component of the article. A number of important data entries have not been treated satisfactorily in Setswana dictionaries, for example; definitions, translations, sense markers, etymology, to mention but a few. The lexicographer has to decide on a selection of entries to treat in the microstructure of the dictionary.

Atkins et al. 1997 give a schematic presentation of typical data types given in comprehensive monolingual and bilingual dictionaries, cf. Tables 27 and 28.

Table 27: Example of data types in a comprehensive monolingual dictionary



1. headword
2. syllabification marks
3. pronunciation
4. variant form
5. inflected form
6. entry subdivision counter
7. entry subdivision
8. part-of-speech label
9. complementation label
10. linguistic label

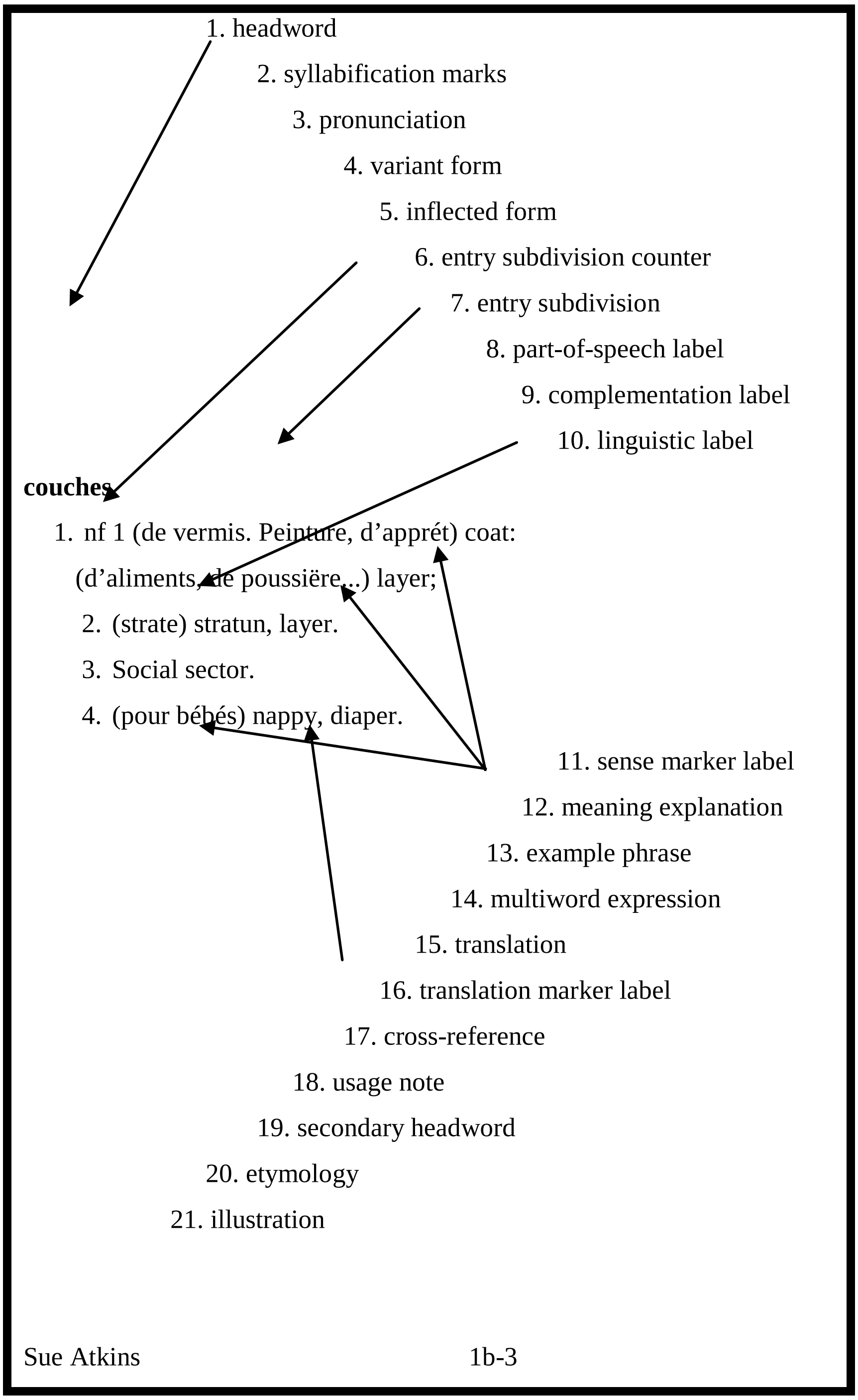
Dread /*dred*/ *v* to fear th greatly: [Vn, V.*ing*] *dread illness, being ill* [V.n *ing*] *I dread my parents finding out*. [V.*to ini*] *We all dread to think what will happen if the factory closes*. [also V.*that*] **dreaded** (also *fml* **dread**) *adj* [*attrib*] greatly feared: *a dreaded disease*. (*joc*) *I heard the dreaded word ‘homework’*.

11. sense marker label
12. meaning explanation
13. example phrase
14. multiword expression
15. translation
16. translation marker label
17. cross-reference
18. usage note
19. secondary headword
20. etymology
21. illustration

Sue Atkins 1b-2

(Monolingual dictionary data type: Atkins et al. 1997)

Table 28: Example of data types in a comprehensive bilingual dictionary



1. headword
2. syllabification marks
3. pronunciation
4. variant form
5. inflected form
6. entry subdivision counter
7. entry subdivision
8. part-of-speech label
9. complementation label
10. linguistic label

couches

1. nf 1 (de vermis. Peinture, d'apprét) coat:
(d'aliments, de poussiëre...) layer;
2. (strate) stratun, layer.
3. Social sector.
4. (pour bébés) nappy, diaper.

11. sense marker label
12. meaning explanation
13. example phrase
14. multiword expression
15. translation
16. translation marker label
17. cross-reference
18. usage note
19. secondary headword
20. etymology
21. illustration

Sue Atkins 1b-3

(Bilingual dictionary data type: Atkins et al. 1997)

5.3 Corpora as a key to writing better dictionary articles

5.3.1 Corpora as an aid to sense distinctions

According to Prinsloo and De Schryver (2004:4), the lexicographer is always in doubt whether he or she has covered all the relevant senses of a lemma sign in the definition (also called paraphrase of meaning) or translation equivalent paradigm. In terms of Prinsloo and Gouws (1996:43), corpus lines will assist the lexicographer in respect of sense distinction, deciding on translation equivalents, retrieval of typical collocations, pinpointing frequent clusters and the selection of representative, authentic examples to be included in the dictionary.

Prinsloo and De Schryver further state that the lexicographer should be cautious not to regard each corpus line as a different sense but rather learn to ‘see the senses emerge’ from a digestible number of corpus lines studied. It is normally also not the intention to study thousands of corpus lines for each lemma but rather to look at a few hundred lines sorted in this sensible ways, e.g. on the word preceding/following the lemma. These lines help the lexicographer to distinguish various senses of the word. The chances of a dictionary compiler gathering all senses and sub-senses on the basis of intuition are zero. Consider the corpus lines in Table 29 which is an extract from the South African Setswana corpus.

Table 29: Concordance lines for *thari* (210) in the South African Setswana corpus

Mateo wee! maitseo a ile kae? Bana ba	***thari***	e ntsho, Bana na mmlala wa sebilo,
ngwe- nyana a mmolelela gore gatwe	***thari***	ya ngwana e kwa ga-bona-mogolo
go feta baabo abo a ilwa o kabo belege	***thari***	wa itlhoboga aw ikela le naga ja
Morongwenyana a raya Morongwe a re:	***thari***	ya ngwana e kae?” Morongwe a gama
Bana ba Afrerika ka bopara, Bana ba	***thari***	e ntsho dinatla. Ke bone maloba
sadi ba baswa le ba ba santseng ba bona	***thari***	Mo moletlong wa lenyalo kgotsa

Mokgalajwe le batho ba Madibe ba fitlha	***thari***	kwa phitlhong mo- rago ga go t
Shole o ntete. Bobedi re ke re supe	***thari***	mmogo, Re supe fa kgole e e le
g, gonne ke fano Modimo o mo timile	***thari***	Jaanong …” “A re tlogeleng

The following senses clearly emerge from the concordance lines listed in Table 29:

1. (late coming)

*Goroga ka nako mo tirong. Ga ke a tla **thari**. A gakologelwa gape….*

(Arrive on time at work. I am not late. He remembered

2. (skin used to carry a baby)

*…e ke Mmabatho mmarona, sebelega bana ka **thari** mpeng. Ka yona e kete nka be ke le thata ka tsaya **thari** ka go belega*

(She is our real mother who always looks after her children. I wish I was strong enough to carry you with *thari*)

3. (black nation) *thari e ntsho*

*…ee Batswana a re tswaneng re le bana ba **thari** e ntsho. Modisetshaba o re boloke*

(…yes, let’s unite as a **black nation**. Foreigners to care for us)

4. (giving birth)

*Modimo o ne o sa tima Motlalepule **thari**, le ene o ka bo a bua monate*

(God did not deprive Motlalepule the opportunity to bear **children**, she should also be proud of that fact)

Table 30: Concordance lines for *tshela* (1540) in the South African Setswana corpus

16) A fitlha a itulela gone. (17) A	***tshela***	ka tsie le dinotshe, (18) Le
go tlhola ka a ne a se mo ntlong. Go	***tshela***	bofofu ntlheng ya lesaka a du
oga le rre. Ke eletsa gore a ka be a sa	***tshela***	gore a tle a bone maungo a lo
a bothito a ithobe dingalo. Mmatshepe a	***tshela***	metse, a be a a loka ka letsw
fetsa matsatsi nae, a mpotsa le gore o	***tshela***	jang. Maswe a diatla. Morago a
magalapa a morubisi; Losika lwa bo-	***tshela***	-le-baloi. Ntomolele a gana nn
Madiba ano a tletse dikwena tea bo-	***tshela***	-ke-go-garume. A tletse marara

The following senses clearly emerge from the concordance lines listed in Table 30:

1. (survival, to live on)

A fitlha a itulela gone. A **tshela** *ka tsie le dinotshe*

(He arrived and settled there. He **survived** on locusts and bees)

2. (alive)

Olga *le rre. Ke eletsa gore a ka be a sa* ***tshela*** *gore a tle a bone maungo*

(Olga and her father. I wish he was still **alive** to witness the outcomes)

3. (cross)

*Madibana a tletse dikwena tla bo-****tshela****- ke-go- garume. A tletse marara*

(The crocodiles are lying in wait to **cross** so that they may attack)

4. (pour)

…a bothito a ithobe dingalo. Mmatshepe a ***tshela*** *metse, a be a a loka ka letswai*

(…apply warm water to ease the pain. Mmatshepe **pours** water and included salt in)

Corpora can furthermore assist the lexicographer in finding typical collocations and combinations of words as computed with WordSmith Tools.

5.3.2 Corpora as an aid to retrieve typical collocations

In this section we will illustrate how a detailed analysis of corpus lines, in combination with frequency counts at various levels, for the frequent Setswana synonyms *batla* and *senka* (to look for) enables lexicographers to enhance the quality of microstructural elements. The aim here is to find the means of getting the relationships between *batla* and *senka* in terms of various statistics generated by WordSmith Tools as shown in Tables 31 and 34 below.

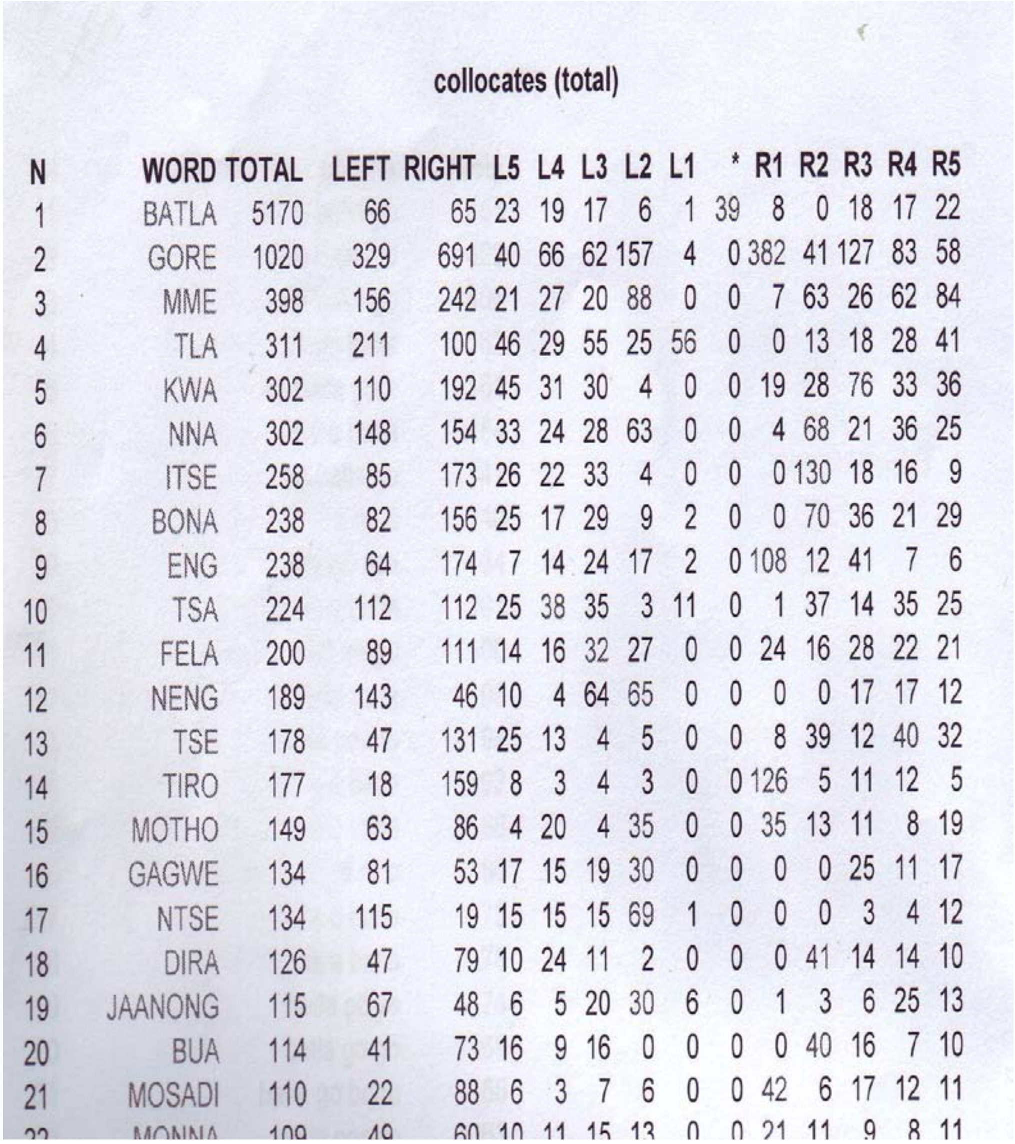
5.3.2.1 Collocates of *batla* according to the South African Setswana corpus

Table 31: Corpus lines for *batla* (5170) in the South African Setswana corpus

. Fa o pota ka fa o utlwe go twe re	***batla***	motho yo o falo- tseng materik
a a tlhole a tshela. o setse a sule! Ke	***batla***	go mo utlwisa seo a ntseng a se la
ela pelo. "Mme o raya jang? Kana ke	***batla***	go ya sekolong," a bua a hupe-
aana ka ena, mma. A re o bone io, mme o	***batla***	go ikgolega ka ena. Re romilwe ke
, 0 didimaletseng? Motswasele: Kana o	***batla***	ke bua eng? Modise: Batho ba tsa
ona. Yo o mpatlang, o tshwanetse go	***batla***	fa moraka wa kgomo o leng ten"
lhogo. (2) (c) Leubajaaka motho, le	***batla***	eng mo morafeng? (3) (d) Tiris
"Ao! " Maipelo a	***batla***	a tshwara sengwenyana mo sefatlheg
panyi, jk. 44 Tumelo Kganetso Mme o	***batla***	gore ke nne le mosa- Mme o batla g
i Dikwalo ke di	***batla***	a di huparetse mo legwafeng. K
diphatsa lwa bone. A ithaya a re o	***batla***	go sala mo polasing fa ba tsamaya.
Nyaa tsala, wena a re ye gae. Ke	***batla***	gore o wele makgwafo. Ke bona
atshe, tshwene. Motho a fosa a ba a	***batla***	a tlhoma ka nko moseja ole. Ya re

The top ten collocates of *batla* occurring immediately to the right of *batla* in the South African Setswana corpus are shown in Table 32 namely *gore* (so that), *mme* (but), tla (will), *kwa* (there), *nna* (me), *itse* (know), *bona* (they), *eng* (what) and *tsa* (for).

Table 32: Collocates of *batla* generated by WordSmith Tools



collocates (total)

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	BATLA	5170	66	65	23	19	17	6	1	39	8	0	18	17	22
2	GORE	1020	329	691	40	66	62	157	4	0	382	41	127	83	58
3	MME	398	156	242	21	27	20	88	0	0	7	63	26	62	84
4	TLA	311	211	100	46	29	55	25	56	0	0	13	18	28	41
5	KWA	302	110	192	45	31	30	4	0	0	19	28	76	33	36
6	NNA	302	148	154	33	24	28	63	0	0	4	68	21	36	25
7	ITSE	258	85	173	26	22	33	4	0	0	0	130	18	16	9
8	BONA	238	82	156	25	17	29	9	2	0	0	70	36	21	29
9	ENG	238	64	174	7	14	24	17	2	0	108	12	41	7	6
10	TSA	224	112	112	25	38	35	3	11	0	1	37	14	35	25
11	FELA	200	89	111	14	16	32	27	0	0	24	16	28	22	21
12	NENG	189	143	46	10	4	64	65	0	0	0	0	17	17	12
13	TSE	178	47	131	25	13	4	5	0	0	8	39	12	40	32
14	TIRO	177	18	159	8	3	4	3	0	0	126	5	11	12	5
15	MOTHO	149	63	86	4	20	4	35	0	0	35	13	11	8	19
16	GAGWE	134	81	53	17	15	19	30	0	0	0	0	25	11	17
17	NTSE	134	115	19	15	15	15	69	1	0	0	0	3	4	12
18	DIRA	126	47	79	10	24	11	2	0	0	0	41	14	14	10
19	JAANONG	115	67	48	6	5	20	30	6	0	1	3	6	25	13
20	BUA	114	41	73	16	9	16	0	0	0	0	40	16	7	10
21	MOSADI	110	22	88	6	3	7	6	0	0	42	6	17	12	11
22	MONNA	109	49	60	10	11	15	13	0	0	21	11	9	8	11

According to Leech (1981:17), collocates consists of the associations of meaning in a particular environment. If one instructs the corpus query tool to calculate and list collocates of the verb *batla*, certain useful conclusions can be drawn. For example, the frequent use of *gore* (that), *kwa* (at), *eng* (what), *fela* (just), *nna* (me), *itse* (know) etc which otherwise escaped attention in dictionaries compiled on intuition.

The concordance lines in WordSmith Tools allows one to see the items that are most frequently found to the left and to the right of a search-word as reflected in Table 32. In Table 32 '*gore*' (so that) collocates 386 times with *batla* in the horizon L1-R1. 4 of these collocates occur to the left of *batla*, 382 to the right. The breakdown of occurrences to the right is 382 times R1, 41 times R2, 127 times R3, 83 times R4 and 58 times R5. The second most frequent collocate of *batla* is the word *eng* (what) which collocates 110 times in the horizon L1-R1. 2 of these collocates occur to the left of *batla* and 108 to the right.

Table 33: Top ten collocates of the base *batla* that collocate immediately to the right of *batla* in the South African Setswana corpus

Base + Collocate	Translation	Frequency
1. *O* ***batla*** *gore*	(wants to)	382
2. *A* ***batla*** *kwa*	(seek there)	19
3. *Ke* ***batla*** *eng*	(looking for)	108
4. *Ba* ***batla*** *motho*	(look for someone)	35
5. ***Batla*** *fela*	(just look)	24
6. *Ne a* ***batla*** *tse*	(wanted to come)	8
7. *Ba* ***batla*** *tiro*	(seek for job)	126
8. ***Batla*** *mme*	(want to know)	7
9. *Fa o* ***batla*** *mosadi*	(looking for a woman)	42
10. *O* ***batla*** *monna*	(looking for a man)	21

It is important to note that there is a recurrent pattern as far as words following *batla* are concerned as indicated in Table 33. One can see demonstratives, conjunctions and the subjectival concord *o* in *o batla go* (he wants to) with the highest frequency followed by *batla gore* (look for), *batla kwa* (look there), *batla eng* (look for what), *batla fela* (look for just) etc, with the lowest frequency. The following is a second example of the corpus lines for the collocates of *senka*.

5.3.2.2 Collocates of *senka* according to the South African Setswana corpus

Table 34: Corpus lines for *senka* (244) in the South African Setswana corpus

e ile." Motsei a re, "A ga ba ka ba	***senka***	ngaka e nngwe gape?" Bikibiki
o a rona a ne a tletsetletse naga re lo	***senka***	mme re sa bone sepe. Phaladi o bil
ke diolo a ntse a apaapa ka dinao a	***senka***	ka tlhoafalo mo lefifing la bosigo
lhelela Keabetswe a ithwala a ya go	***senka***	tiro. Aitsane le fa 0 ka~ seatla m
ao nnyaa Ao, ga se nna! apaapa	***senka***	mo fifing ka diatla Ke ntse ke
Nakedi a nna a apaapa mo lefifing mme a	***senka***	mo- kgwaro. E rile fa a sena g
selekanyo sefe? | Ba ya kae? Ba	***senka***	go le kae ? kaBla supetsa; dir
g e re fa boroko bo fedile ke ye go	***senka***	dikala di le pedi pele ke ya go th
phiri sa me, ke tla bo ke ikaletsa. A o	***senka***	gore ke ipolaye ? Ema, ke mo
di tswa mo mometsong. mpampetsa	***senka***	ka bonolo; itaya ka bonolo
lobelo lo logolo ka di ne di gopotse go	***senka***	kwa metsi a ka bonwang gone. 19
kelong jwa dijo. Ga ba kitla ba re	***senka***	gona. Ba tla fisa ntlo. Re tla swe

The following senses clearly emerge from the concordance lines listed in Table 34:

1. (seek)

… *sepe ka go boela Gauteng. O ne a ya go* ***senka*** *tiro kwa Vereeniging, mme a ithuta*

(… nothing by going back to Gauteng. He went to **seek** job in Vereeniging where he studied)

2. (find)

… *atsaya lobone. O ne a apaapa mo lefifing a* ***senka*** *dikgetse tse a di alang*

(… he took the lamp. He gropes in the darkness and **find** the sacks to spread out)

3. (search)

Badisa ba tswa ba ya go ***senka*** *dinku, ba di fitlhela di eme*

(The shepherds went out to **search** for the sheep, they found them waiting)

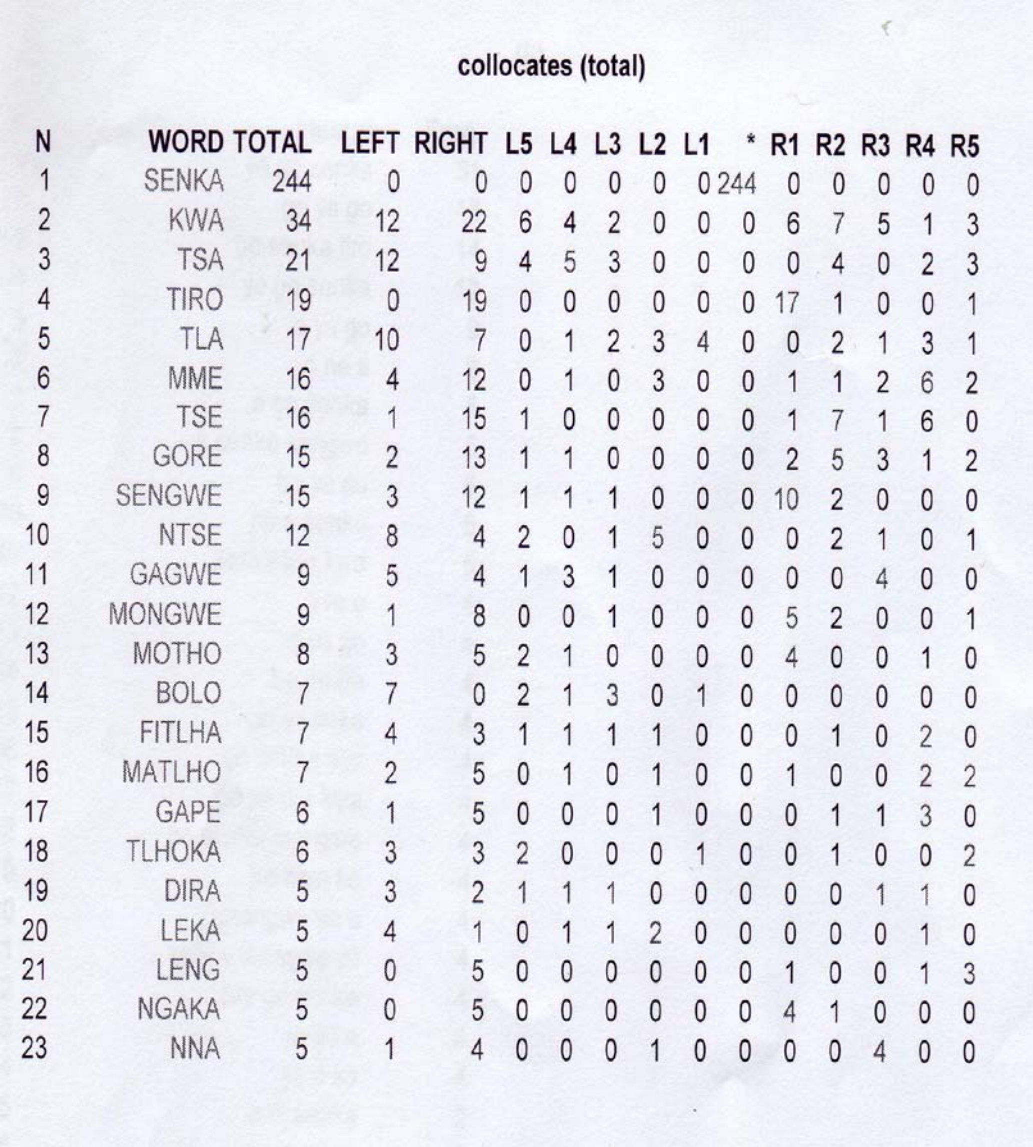
4. (want)

… *fa ba ka se ka ba nthusa; ba re ba* ***senka*** *barutintshi ba ba boitshoko*

(… if they are not going to assist me; they **want** patient teachers)

Consider the following collocates generated by the WordSmith Tools to identify the co-occurrence to the right and left of *senka*:

Table 35: Collocates of *senka* generated by the WordSmith Tools



collocates (total)

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	SENKA	244	0	0	0	0	0	0	0	244	0	0	0	0	0
2	KWA	34	12	22	6	4	2	0	0	0	6	7	5	1	3
3	TSA	21	12	9	4	5	3	0	0	0	0	4	0	2	3
4	TIRO	19	0	19	0	0	0	0	0	0	17	1	0	0	1
5	TLA	17	10	7	0	1	2	3	4	0	0	2	1	3	1
6	MME	16	4	12	0	1	0	3	0	0	1	1	2	6	2
7	TSE	16	1	15	1	0	0	0	0	0	1	7	1	6	0
8	GORE	15	2	13	1	1	0	0	0	0	2	5	3	1	2
9	SENGWE	15	3	12	1	1	1	0	0	0	10	2	0	0	0
10	NTSE	12	8	4	2	0	1	5	0	0	0	2	1	0	1
11	GAGWE	9	5	4	1	3	1	0	0	0	0	0	4	0	0
12	MONGWE	9	1	8	0	0	1	0	0	0	5	2	0	0	1
13	MOTHO	8	3	5	2	1	0	0	0	0	4	0	0	1	0
14	BOLO	7	7	0	2	1	3	0	1	0	0	0	0	0	0
15	FITLHA	7	4	3	1	1	1	1	0	0	0	1	0	2	0
16	MATLHO	7	2	5	0	1	0	1	0	0	1	0	0	2	2
17	GAPE	6	1	5	0	0	0	1	0	0	0	1	1	3	0
18	TLHOKA	6	3	3	2	0	0	0	1	0	0	1	0	0	2
19	DIRA	5	3	2	1	1	1	0	0	0	0	0	1	1	0
20	LEKA	5	4	1	0	1	1	2	0	0	0	0	0	1	0
21	LENG	5	0	5	0	0	0	0	0	0	1	0	0	1	3
22	NGAKA	5	0	5	0	0	0	0	0	0	4	1	0	0	0
23	NNA	5	1	4	0	0	0	1	0	0	0	0	4	0	0

In Table 35 item 4, *tiro* (work) collocates 17 times with *senka* in the horizon L1-R1. 17 of these occur to the right and 0 to the left. The breakdown of occurrences to the right is 17 times R1, 1 R2, 0 R3, 0 R4 and 1 R5. When item 8 *gore* (so that) is taken, it collocates 2 times with *senka* in the horizon L1-R1. 2 of these collocates occur 2 times to the right *of senka* and 0 times to the left. The low frequent use of *senka gore* is because the form is more widely used in Botswana than in South Africa.

Table 36: Top ten collocates of the base *senka* that occur immediately to the right of *senka* in the South African Setswana corpus

Base + Collocate	**Translation**	**Frequency**
1. *Senka leng*	(find when)	1
2. *Senka tiro*	(find the job)	17
3. *Senka gore*	(find that)	2
4. *Senka motho*	(find a person)	4
5. *Senka tse*	(find this)	1
6. *Senka sengwe*	(find something)	10
7. *Senka mme*	(find a mother)	1
8. *Senka kwa*	(find there)	6
9. *Senka mongwe*	(find someone)	5
10. *Senka ngaka*	(find a doctor)	4

Table 37 below will now be used to contrast the presumed Setswana synonyms *batla* and *senka*. When Table 35 and Table 36 are compared using the nodes *gore* and *tiro* to see which items are most frequently found to the right of the search-word *batla* and *senka*, the word '*batla gore*' appears 691 times and *senka gore* appears 13 times. The results suggest that the Setswana users in South Africa are more likely to use *batla* than *senka*.

Table 37: Comparison between *batla* and *senka* in collocation with *gore* and *tiro*

Word	**Left**	**Right**	**Total**
batla gore (want to)	329	691	1020
senka gore (look for)	2	13	15
batla tiro (look for a job)	18	159	177
senka tiro (find a job)	0	19	19

Compare another example which further illustrates the contrast between *batla* and *senka*:

When the first 20 collocates are compared between *batla* and *senka* certain collocates are missing for both the so-called synonyms as indicated by an ‘x’ in Table 38.

Table 38: Comparison of the top ten collocates of *batla* and *senka*

Collocates	***Batla***	***Senka***
1) *mme*	✓	✓
2) *tla*	✓	✓
3) *kwa*	✓	✓
4) *itse*	✓	x
5) *bona*	✓	x
6) *eng*	✓	x
7) *matlho*	x	✓
8) *sengwe*	x	✓
9) *dira*	x	✓
10) *nna*	x	✓

5.3.3 Corpora as an aid to pinpoint clusters

Clusters are words which are found repeatedly in each other's company. There is a recurrent pattern for the words following *batla* and *senka*. In this case the lexicographer can instruct the corpus query tool to calculate and list the frequent clusters. Consider the following:

Table 39: Two-word clusters with *batla* (451) in the initial position in South African Setswana corpus

5170 Cluster		**Frequency**	**Percentage**
batla	*gore*	1020	19.72 %
batla	*mme*	398	7.69 %
batla	*tla*	311	6.015 %
batla	*kwa*	302	5.84 %
batla	*nna*	302	5.84 %
batla	*itse*	258	4.99 %
batla	*bona*	238	4.60 %
batla	*eng?*	238	4.60 %
batla	*tsa*	224	4.33 %

Going down one level in respect of words in combination with *batla* as indicated in Table 39, *batla* is followed 1020 times by *gore*, 19.72% of all occurrences of *batla* and by *mme* 398, 7.69% of all occurrences of *batla* and a variety of other clusters ranging from 6,015% to 4, 33%. This means that *batla gore* should definitely be considered for inclusion in the article of *batla* in any Setswana dictionary, which is not the case in existing dictionaries.

Table 40: Three word clusters with *batla* (451) in the initial position in the South African Setswana corpus

451 Cluster	**Frequency**	**Percentage**
a batla go	422	93.6 %
ke batla go	406	90.02 %
ya go batla	163	36.14 %
ke batla gore	161	35.7 %
ne a batla	154	34.15 %
ba batla go	141	31.26 %
batla go itse	134	29.71 %
o batla gore	98	21.72 %

From Table 40 moving downwards with words used in combination from left to right, one can see that *batla* appears 422 times between a ~ go and 406 times between ke ~ go. It is important to note that *batla go* preceded by *a* counts 422 or 93,6% out of the possible 451 occurrences of *batla* and *batla go* preceded by *ke* counts 406 or 90,02% of the possible 451 of the occurrences of *batla* with the rest ranging from 6,14% to 21,72%. This means that inclusion of the subject concord *ke* and or *a* as well as the infinitive prefix *go* in the article of *batla* is highly recommended.

It is therefore important for lexicographers to examine synonyms thoroughly before giving the translation equivalents and definitions including cross-reference entries such as *batla* and *senka*.

It is clear that the examples in concordance lines in Table 33 and Table 36 bring about the difference between *batla* and *senka*. True synonyms are rare in Setswana. It is a matter of dialectical preference. For example, *senka* is preferred by the Southern Setswana sub-group, *Setlhaping* and *Setlharo* dialects in the Taung, Vryburg and Kuruman district. *Batla* is preferred by the Eastern Setswana which comprises of the *Sekwena* and *Sekgatla* sub-dialects spoken around Pretoria.

In the next section we will highlight problems relating to the treatment of verbs, polysemy and synonyms in the existing Setswana dictionaries and the Setswana months will be critically analyzed against the background information of the English and the Afrikaans dictionaries.

5.3.4 Corpora as an aid to select typical and natural examples

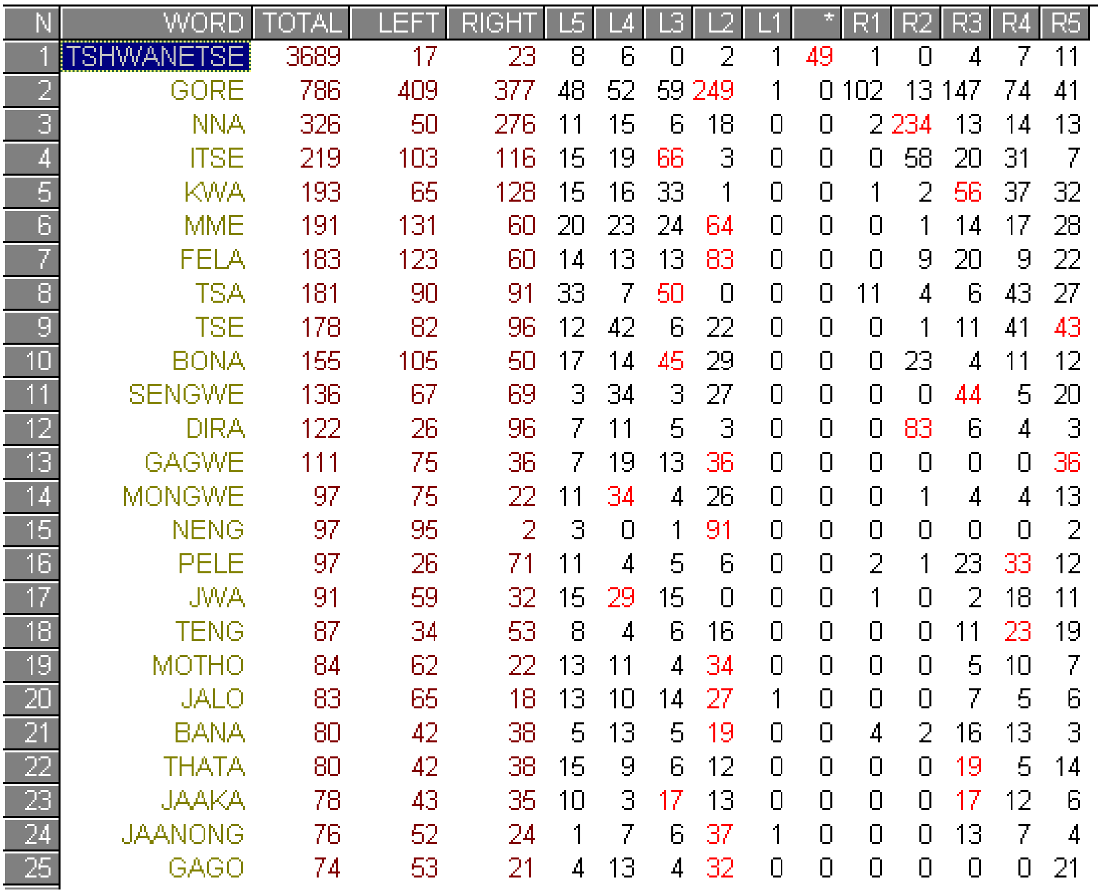
In this section we will look into the huge potential of combining different corpus query tools, with special reference to the selection of excellent typical and natural examples. According to Fox (1987:138), the terms typical and natural examples can be defined as follows:

> "Our first and foremost requirement for examples is typicality: that they should show the way in which people actually use the word they are examplifying. […] naturalness […] is the well-formedness of sentences not in isolation but in text."

Laufer (1992:72) also stated that lexicographers who are educated native speakers of the language are bound to have correct intuitions about their mother tongue, about the grammaticality of the word, its typical use and its typical environment. These intuitions are necessarily less correct than intuitions of those language users who are represented in the corpus and are therefore not less reliable.

The lexicographer can thus combine the output of different good query tools such as word-frequency counts and concordance line screens. For example,

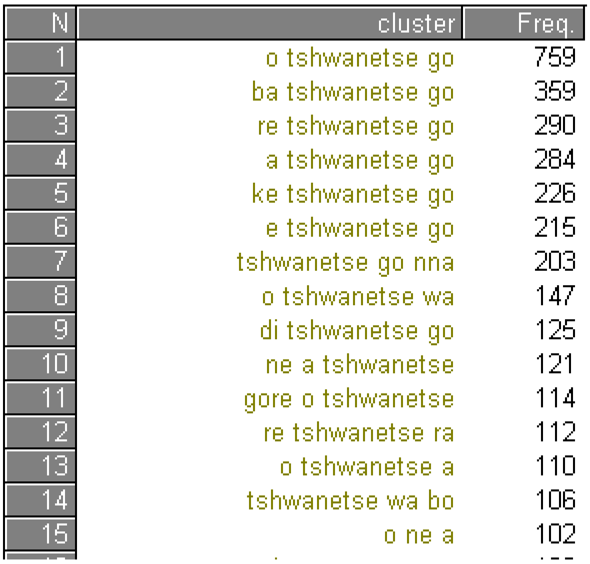
Table 41: Collocates of the base *tshwanetse* (with horizons L5-R5) in the South African Setswana corpus



N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	TSHWANETSE	3689	17	23	8	6	0	2	1	49	1	0	4	7	11
2	GORE	786	409	377	48	52	59	249	1	0	102	13	147	74	41
3	NNA	326	50	276	11	15	6	18	0	0	2	234	13	14	13
4	ITSE	219	103	116	15	19	66	3	0	0	0	58	20	31	7
5	KWA	193	65	128	15	16	33	1	0	0	1	2	56	37	32
6	MME	191	131	60	20	23	24	64	0	0	0	1	14	17	28
7	FELA	183	123	60	14	13	13	83	0	0	0	9	20	9	22
8	TSA	181	90	91	33	7	50	0	0	0	11	4	6	43	27
9	TSE	178	82	96	12	42	6	22	0	0	0	1	11	41	43
10	BONA	155	105	50	17	14	45	29	0	0	0	23	4	11	12
11	SENGWE	136	67	69	3	34	3	27	0	0	0	0	44	5	20
12	DIRA	122	26	96	7	11	5	3	0	0	0	83	6	4	3
13	GAGWE	111	75	36	7	19	13	36	0	0	0	0	0	0	36
14	MONGWE	97	75	22	11	34	4	26	0	0	0	1	4	4	13
15	NENG	97	95	2	3	0	1	91	0	0	0	0	0	0	2
16	PELE	97	26	71	11	4	5	6	0	0	2	1	23	33	12
17	JWA	91	59	32	15	29	15	0	0	0	1	0	2	18	11
18	TENG	87	34	53	8	4	6	16	0	0	0	0	11	23	19
19	MOTHO	84	62	22	13	11	4	34	0	0	0	0	5	10	7
20	JALO	83	65	18	13	10	14	27	1	0	0	0	7	5	6
21	BANA	80	42	38	5	13	5	19	0	0	4	2	16	13	3
22	THATA	80	42	38	15	9	6	12	0	0	0	0	19	5	14
23	JAAKA	78	43	35	10	3	17	13	0	0	0	0	17	12	6
24	JAANONG	76	52	24	1	7	6	37	1	0	0	0	13	7	4
25	GAGO	74	53	21	4	13	4	32	0	0	0	0	0	0	21

In Table 41, a selection of the collocates of the base *tshwanetse* with the horizon L5-R5 is listed. In Table 41 items *gore* (so that) collocates 103 times with *tshwanetse* in the horizon of L1-R1. 1 of these collocates occur on the left of *tshwanetse*, 102 to the right. The second most frequent collocate of *tshwanetse* is *nna* (me) which collocates 252 times in the horizon L2-R2. 18 of these collocates occur on the left of *tshwanetse* and 234 to the right.

Table 42: Three-word clusters with tshwanetse (909) in the South African Setswana corpus



N	cluster	Freq.
1	o tshwanetse go	759
2	ba tshwanetse go	359
3	re tshwanetse go	290
4	a tshwanetse go	284
5	ke tshwanetse go	226
6	e tshwanetse go	215
7	tshwanetse go nna	203
8	o tshwanetse wa	147
9	di tshwanetse go	125
10	ne a tshwanetse	121
11	gore o tshwanetse	114
12	re tshwanetse ra	112
13	o tshwanetse a	110
14	tshwanetse wa bo	106
15	o ne a	102

From Table 42 above we see that the most frequent three-word cluster with *tshwanetse* in the South African Setswana corpus is *o tshwanetse go* (he/she is suppose to). The second most frequent cluster is *ba tshwanetse go* (they are suppose to); followed by *re tshwanetse go* (we are suppose to).

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

5.4 Problems related to the treatment of the Setswana verbs

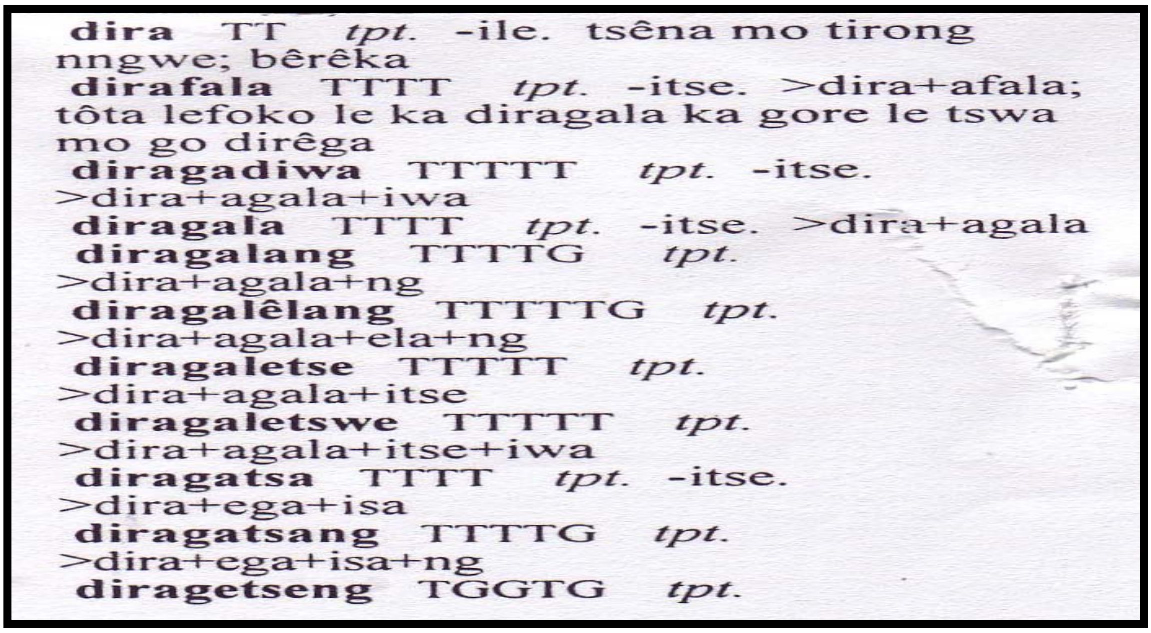
Verbs in Setswana change their forms in order to express, or help express, different perspectives for viewing an action or state, such as the time an event happened, how long it lasted, and the number of different semantic connotations as given for the verb *reka* in Table 22. The base-form focuses on the meaning of a lexical verb without considering its derivations. These verb forms make available some important differences of meanings as reflected in example 35**.** Consider the following derived forms of the verb *dira* (work, do).

Example 35

Applied verb form,	*-tsa*	>	*e.g. diragatsa*	(do something for)
Causative verb form,	*-isa, -ya*	>	*e.g. dirisa*	(cause/let something happen)
Neuter verb form,	*-ela*	>	*e.g. direla*	(work for)
Passive verb form,	*-wa, iwa*	>	*e.g. dirwa*	(be doing something)
Perfective verb form,	*-ile*	>	*e.g. dirile*	(done)
Reciprocal verb form,	*-na*	>	*e.g. dirisana*	(work together)
Reversive verb form,	*-ola, -olola*	>	*e.g. dirolola*	(to undo)

Neither of these forms are clearly defined in the two monolingual Setswana dictionaries as reflected below in the THAN and the THAND. Examples 36 and 37 below are now used to illustrate problems related to the treatment of the verb *'dira'* and its derivatives.

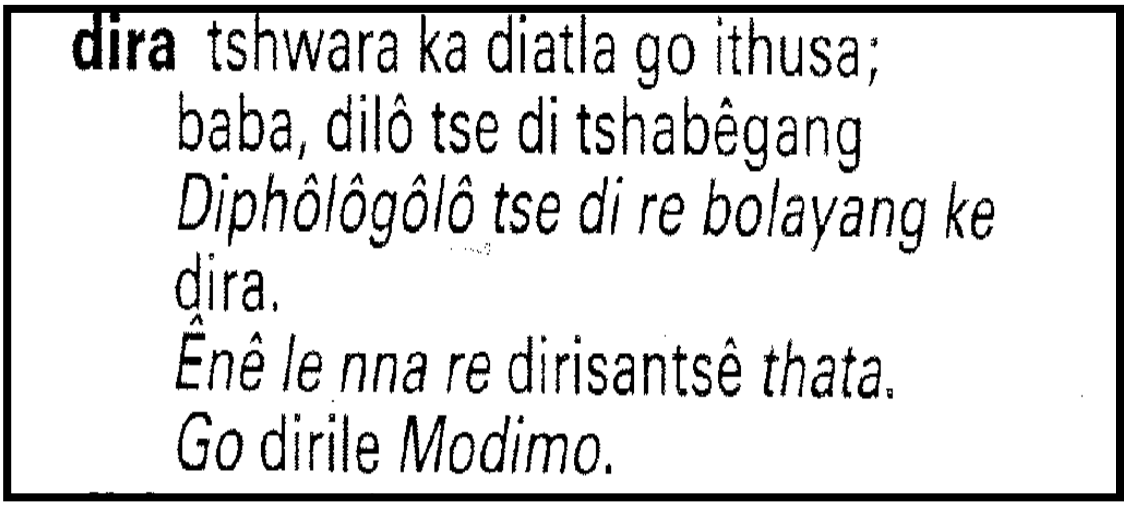
Example 36: THAN



dira TT *tpt.* -ile. tsêna mo tirong
nngwe; bêrêka
dirafala TTTT *tpt.* -itse. >dira+afala;
tôta lefoko le ka diragala ka gore le tswa
mo go dirêga
diragadiwa TTTTT *tpt.* -itse.
>dira+agala+iwa
diragala TTTT *tpt.* -itse. >dira+agala
diragalang TTTTG *tpt.*
>dira+agala+ng
diragalêlang TTTTTG *tpt.*
>dira+agala+ela+ng
diragaletse TTTTT *tpt.*
>dira+agala+itse
diragaletswe TTTTT *tpt.*
>dira+agala+itse+iwa
diragatsa TTTT *tpt.* -itse.
>dira+ega+isa
diragatsang TTTTG *tpt.*
>dira+ega+isa+ng
diragetseng TGGTG *tpt.*

Monolingual dictionaries deal with defining equivalents. However, this is not the case with the THAN as indicated in Example 36 above. The dictionary is pilled up by morphological and grammatical information. No paraphrase of meaning and examples usage are given. It is important for the lexicographer to pay attention to the subject of explaining and not to giving the morphological form only. Finding the meaning of a word is the primary aim of the Setswana dictionary use.

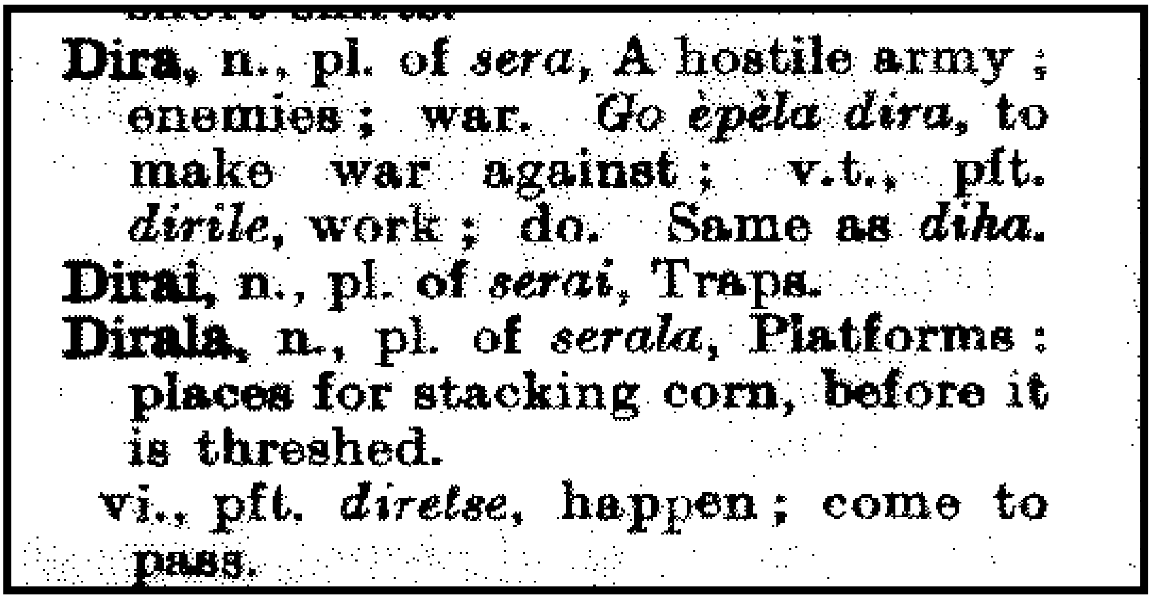
Example 37: THAND



dira tshwara ka diatla go ithusa;
baba, dilô tse di tshabêgang
Diphôlôgôlô tse di re bolayang ke
dira.
Ênê le nna re dirisantsê *thata.*
Go dirile *Modimo.*

Example 37 defines only two senses i.e. ‘enemies’ referring to both wild animals and people and ‘using of hands to work’ while other senses are excluded. Compare in this regard the treatment of the verb *dira* in the following bilingual dictionaries:

Example 38: SED



Dira, n., pl. of *sera*, A hostile army; enemies; war. *Go èpèla dira*, to make war against; v.t., pft. *dirile*, work; do. Same as *diha*.

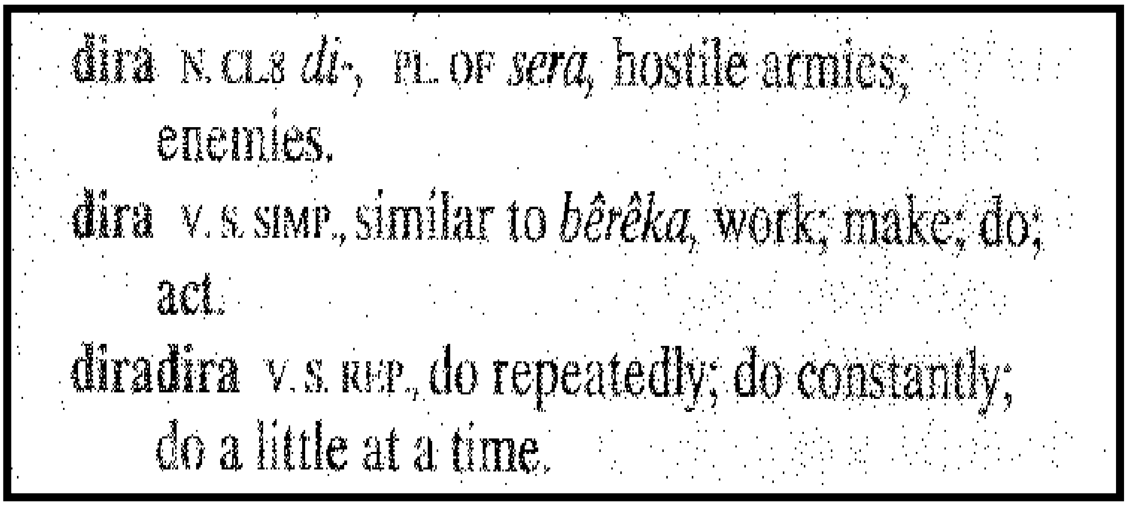
Dirai, n., pl. of *serai*, Traps.

Dirala, n., pl. of *serala*, Platforms: places for stacking corn, before it is threshed.

vi., pft. *diretse*, happen; come to pass.

The compilers in example 38 succeded in giving the translation equivalents of *dira* as enemies, work and do, but often find themselves giving translation equivalents which do not conform to the meaning of the original which often mislead the users. For example, go epela *dira* (to make war against) and pft *dirile* (done).

Example 39: MSED



dira N. CL8 *di-*, PL. OF *sera*, hostile armies; enemies.

dira V. S. SIMP., similar to *bêrêka*, work; make; do; act.

diradira V. S. REP., do repeatedly; do constantly; do a little at a time.

It is clear from the given examples 38 and 39 that more semantic guidance is given compared to the example 37.

In Table 43 below, the article should involve determining the meaning of *dira* in various senses and not only two senses as reflected in examples 37, 38 and 39. Table 43 presents an attempt to improve on typical articles for the verb *dira* and maximally use of corpus data.

Table 43: Corpus lines for *dira* (7074) in the South African Setswana corpus

tlhoka puo. Sy o o logang maano a go	***dira***	bosula o tla bidiwa Rra-bolotsa
ditlhong. Fa o na le tsholofelo, o	***dira***	o sa tetesele. Solofela gonne ga O
O kwena jaaka Banotwa, o ile a ya go	***dira***	legae ja gagwe Mochudi. Lefatshe ja
ladi, monna yo. O nang le maatla go	***dira***	sengwe le sengwe mo Modimolle. Moth
Kana motho yo o latelang kgosi 0 ka	***dira***	dira eng se se iseng se dirwe?i '3Ka
duele ona molao o o re kganelang go	***dira***	O a re tshwara. Bosigo o tshw
a, o mpolelele, ausi .1 Nka tloga ka go	***dira***	dilo!" Mooki a tsamaya ka ntlha y
. . ;- ! ~ 1. (a) Lebota le	***dira***	eng~ (b) Dit~hare di tla nna mo
lhagisiwa: 37 Tumelo Kganetso Ke a	***dira***	ka jalo ke nna le madi Ga ke dire,
e ka mafoko a gago. 4. Thakadu o ne a	***dira***	dira kae? 5. Maikaelelo a ga , mmagwe M
dira, anetse a tlosiwa. Re tshwanetse ra	***dira***	gore Kgomo a mo kobe mo motseng. Re
Jaanong Dafita a ba amogela a ba	***dira***	ditlhogo tsa masomo a batlhaba

From 43 the following senses emerge:

1. (enemies)

Mantsho *ga tlhwe a utlwa sentle.* ***Dira*** *tsa rona di mo boitumelong*

(Mantsho is not yet aware. Our **enemies** are rejoicing)

2. (work)

…a ***dira*** *teng mme a tshwanetse go ya go* **dira** *kwa Taung*. *Pele a tswa ka kgoro*

(he works there, though he should go and **work** in Taung. Before he left)

3. (make)

*…a tse ba di utlwileng. Batho ba **dira** leratla fa ba bona Moatlhodi a ema*

(what they have heard. People **make** noise when the judge stood up

4. (do)

*Fa Ramasedi a rata, a ka e **dira** letsatsi le penne*

(when the almighty God wants, he may **do** it during sun shine)

5.5 The treatment of the Setswana months

The Setswana months are not satisfactorily treated in Setswana monolingual dictionaries. The English and Afrikaans dictionaries will be used to compare the treatment of the Setswana months. Many shortcomings exist as far as the presentation of information is concerned. Entries state only the names of the months in their chronological order, and thus no justice has been done to bring about the historical and cultural aspects. These months are defined inadequately, thus they provide no meaning as illustrated in example 42 below. Two Setswana monolingual dictionaries are compared to English and Afrikaans dictionaries.

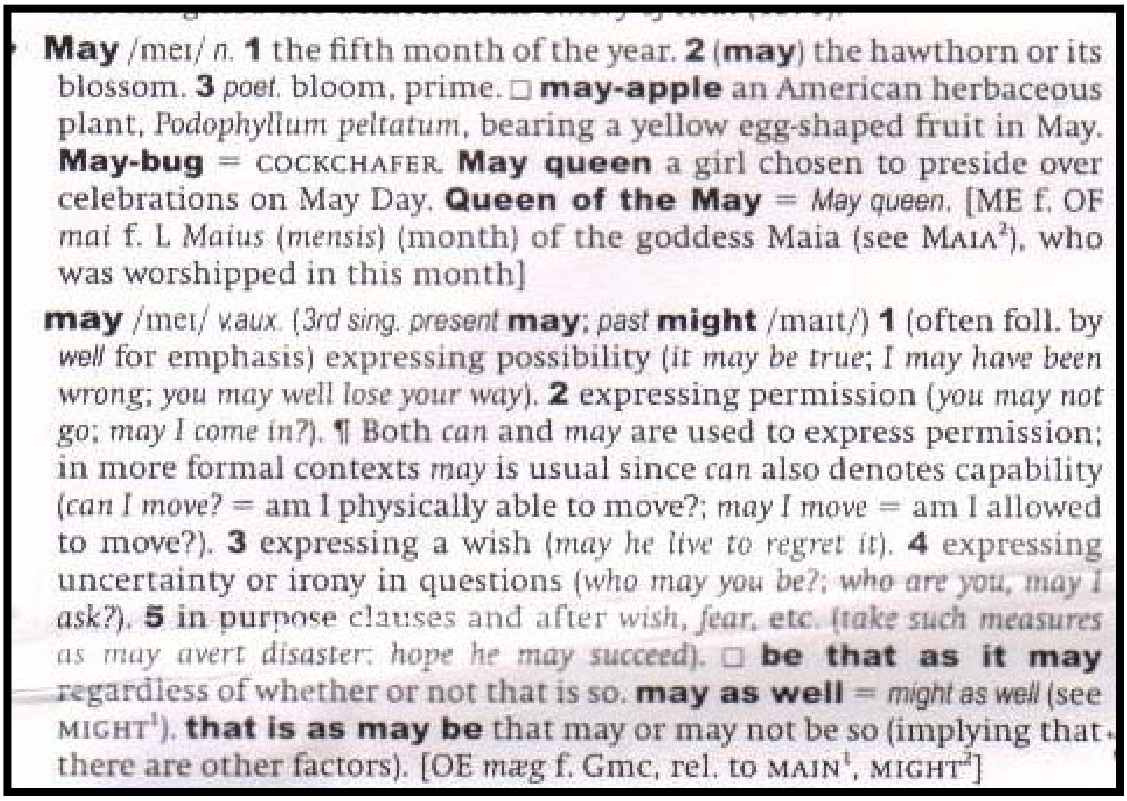
Example 40: THAN and THAND

No.	**THAN**	**THAND**
3	***Mopitlwe****. T.G.T. / mopitlo In / la. bo -. kgwedi ya boraro ya ngwaga.* March. The third month of the year	***3. Mopitlwe*** *kgwedi ya boraro ya ngwaga.* March. The third month of the year
4	***Moranang*** *T.T.T.T. In / la bo – kgwedi ya botlhano ya ngwaga.* April. The fourth month of the year	***4. Moranang*** *kgwedi ya bone ya ngwaga.* April. The fourth month of the year
5	***Motsheganong*** *T.T.T.T. In /la. bo – kgwedi ya botlhano ya ngwaga.* May. The fifth month of the year	***5. Motsheganong*** *kgwedi ya botlhano ya ngwaga e mariga a simologang ka yana.* May. The fifth month of the year, the beginning of winter.

It is unfortunate that the two Setswana monolingual dictionaries have not given the cultural meanings of these lemmas i.e. *Mopitlwe, ngwana wa motswana o ja gore a pipitlelwe ka ntlha fa dijo ele ntletse ntletse.* (The Motswana child eats a lot in such a way that he/she constipate, because there is too much food available).

Aspects of culture are also employed to refer to the aesthetic or intellectual quality of a particular language's art, literature and institution. Many of the Setswana concepts with which we operate are culturally bounded, in the sense that their understanding depends upon socially transmitted knowledge. Compare the article (May month) taken from the Oxford English dictionary (OED) and the Oxford School Dictionary (OSD) where copilers make attempts to cover few examples in bringing about definitions and meaning descriptions.

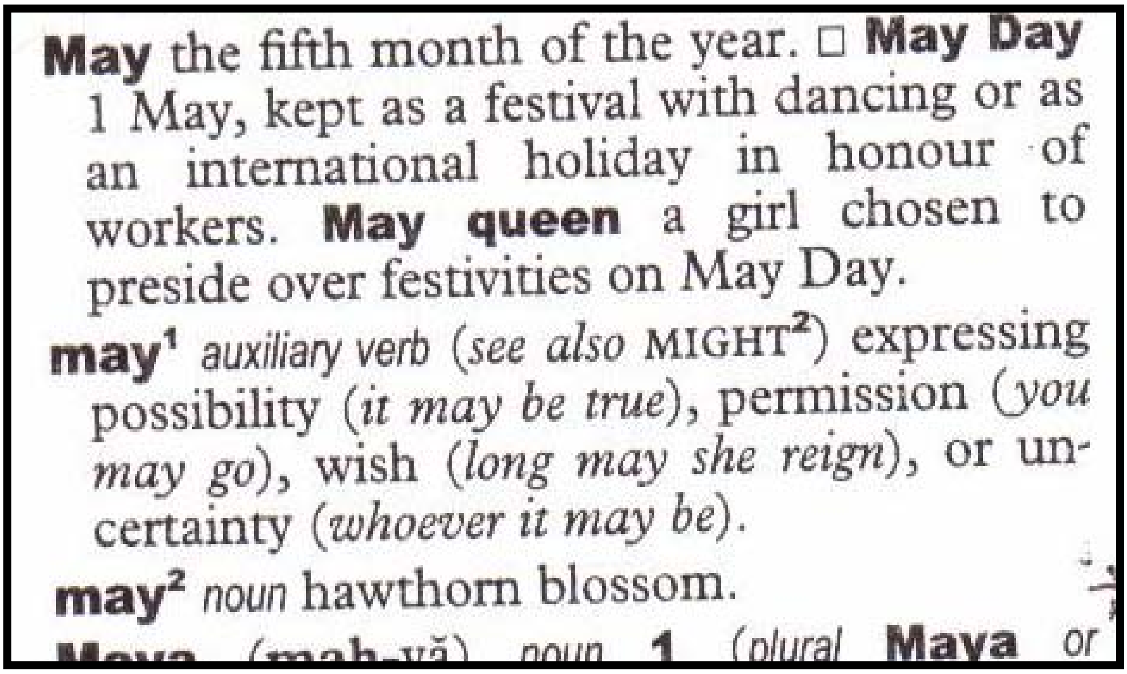
Example 41: OED



May /meɪ/ *n.* **1** the fifth month of the year. **2** (**may**) the hawthorn or its blossom. **3** *poet.* bloom, prime. □ **may-apple** an American herbaceous plant, *Podophyllum peltatum*, bearing a yellow egg-shaped fruit in May. **May-bug** = COCKCHAFER. **May queen** a girl chosen to preside over celebrations on May Day. **Queen of the May** = *May queen*. [ME f. OF *mai* f. L *Maius* (*mensis*) (month) of the goddess Maia (see MAIA^{2}), who was worshipped in this month]

may /meɪ/ *v.aux.* (*3rd sing. present* **may**; *past* **might** /maɪt/) **1** (often foll. by *well* for emphasis) expressing possibility (*it may be true*; *I may have been wrong; you may well lose your way*). **2** expressing permission (*you may not go; may I come in?*). ¶ Both *can* and *may* are used to express permission; in more formal contexts *may* is usual since *can* also denotes capability (*can I move?* = am I physically able to move?; *may I move* = am I allowed to move?). **3** expressing a wish (*may he live to regret it*). **4** expressing uncertainty or irony in questions (*who may you be?; who are you, may I ask?*). **5** in purpose clauses and after *wish, fear,* etc. (*take such measures as may avert disaster; hope he may succeed*). □ **be that as it may** regardless of whether or not that is so. **may as well** = *might as well* (see MIGHT^{1}). **that is as may be** that may or may not be so (implying that there are other factors). [OE *mæg* f. Gmc, rel. to MAIN^{1}, MIGHT^{2}]

Example 42: OSD



May the fifth month of the year. □ **May Day** 1 May, kept as a festival with dancing or as an international holiday in honour of workers. **May queen** a girl chosen to preside over festivities on May Day.

may^{1} *auxiliary verb* (*see also* MIGHT^{2}) expressing possibility (*it may be true*), permission (*you may go*), wish (*long may she reign*), or uncertainty (*whoever it may be*).

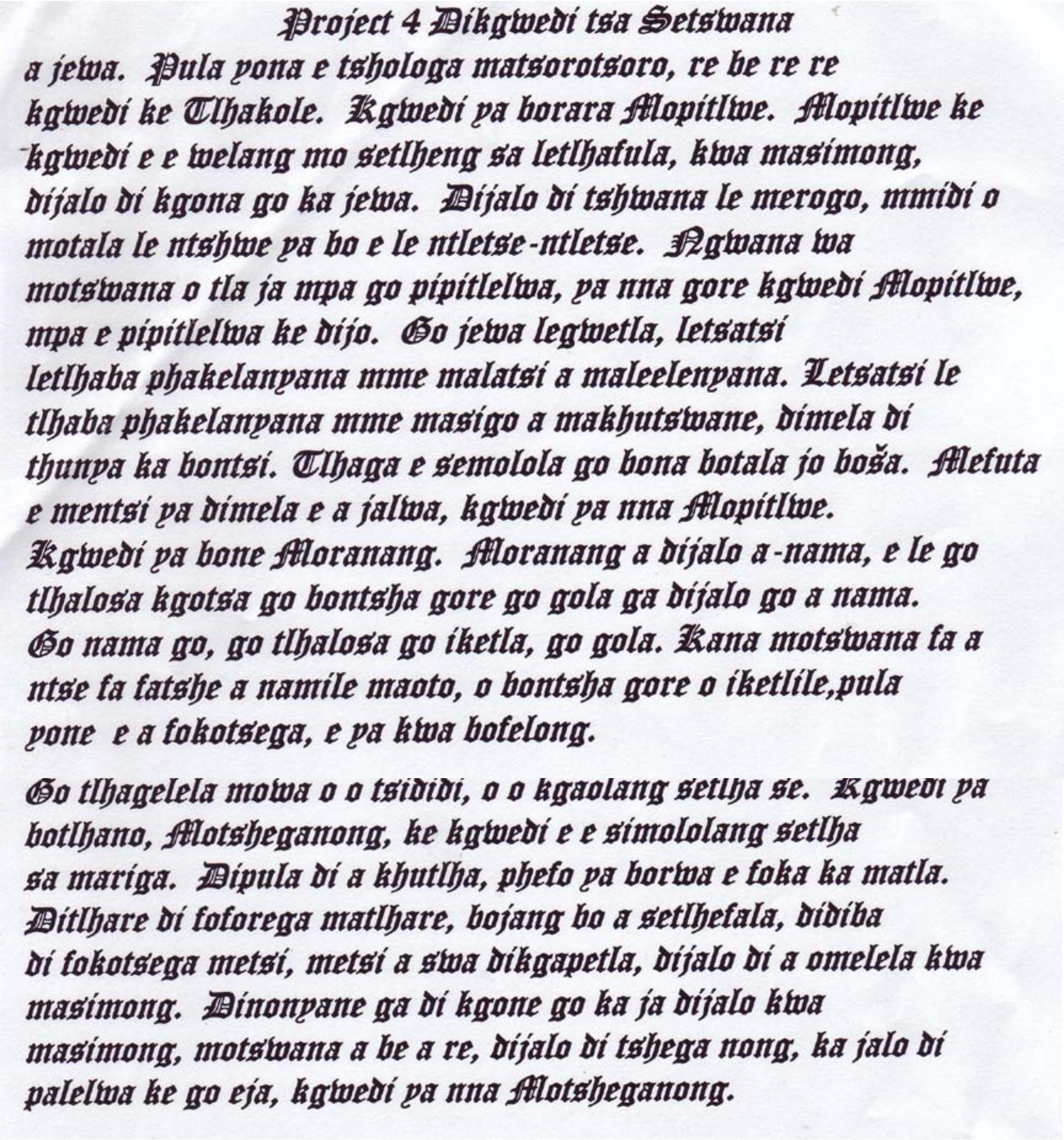
may^{2} *noun* hawthorn blossom.

Maya (mah-yă) *noun* **1** (*plural* **Maya** or

It is important to note that the oral corpus component has brought into light the information that would enrich our dictionary definitions and meaning descriptions.

Consider Table 44 on the Setswana months generated from the South African oral Setswana projects in comparison to the corpus lines generated from the South African written corpus in Table 45.

Table 44: The South African oral Setswana corpus



Project 4 Dikgwedi tsa Setswana

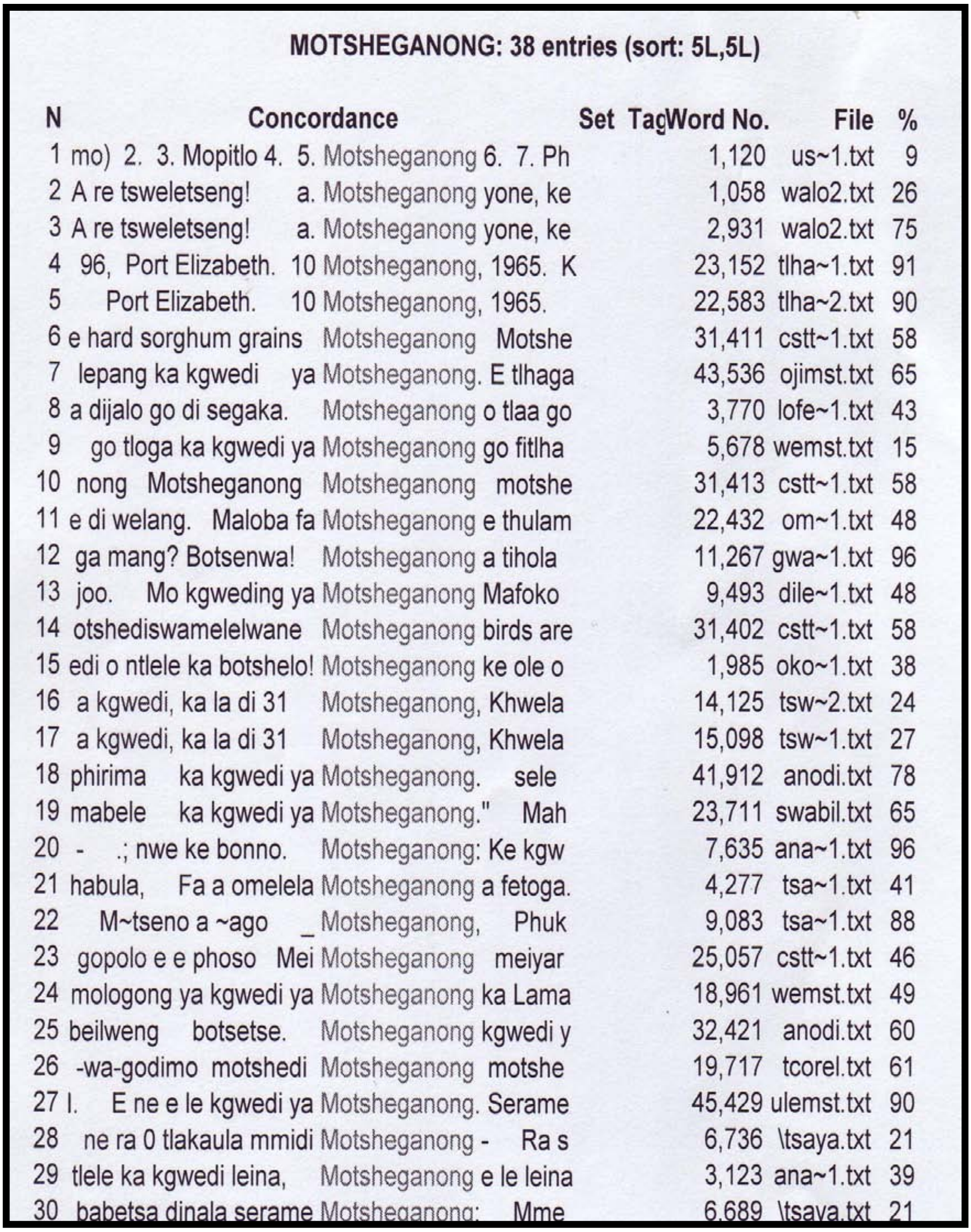
a jewa. Pula yona e tshologa matsorotsoro, re be re re kgwedi ke Tlhakole. Kgwedi ya borara Mopitlwe. Mopitlwe ke kgwedi e e welang mo setlheng sa letlhafula, kwa masimong, dijalo di kgona go ka jewa. Dijalo di tshwana le merogo, mmidi o motala le ntshwe ya bo e le ntletse-ntletse. Ngwana wa motswana o tla ja mpa go pipitlelwa, ya nna gore kgwedi Mopitlwe, mpa e pipitlelwa ke dijo. Go jewa legwetla, letsatsi letlhaba phakelanyana mme malatsi a maleelenyana. Letsatsi le tlhaba phakelanyana mme masigo a makhutswane, dimela di thunya ka bontsi. Tlhaga e semolola go bona botala jo boša. Mefuta e mentsi ya dimela e a jalwa, kgwedi ya nna Mopitlwe. Kgwedi ya bone Moranang. Moranang a dijalo a-nama, e le go tlhalosa kgotsa go bontsha gore go gola ga dijalo go a nama. Go nama go, go tlhalosa go iketla, go gola. Kana motswana fa a ntse fa fatshe a namile maoto, o bontsha gore o iketlile,pula yone e a fokotsega, e ya kwa bofelong.

Go tlhagelela mowa o o tsididi, o o kgaolang setlha se. Kgwedi ya botlhano, Motsheganong, ke kgwedi e e simololang setlha sa mariga. Dipula di a khutlha, phefo ya borwa e foka ka matla. Ditlhare di foforega matlhare, bojang bo a setlhefala, didiba di fokotsega metsi, metsi a swa dikgapetla, dijalo di a omelela kwa masimong. Dinonyane ga di kgone go ka ja dijalo kwa masimong, motswana a be a re, dijalo di tshega nong, ka jalo di palelwa ke go eja, kgwedi ya nna Motsheganong.

Table 44 can now be interpreted as follows:

During this time of the second month of the year, it is heavingly raining, and then the Batswana said: *e tlhakola mogote wa letsatsi*. They enjoy autumn season and the sun rises early and the days are longer and the nights are shorter. The plants and grass are becoming green. Most of the cultivation is taking place during the month of March. As a result the Batswana children are eating over to such an extent that they got constipated. During the month of April, most of the plants are showing their colours. This means that most of the plants are growing well. The Batswana people are relaxing and enjoying and the time of rain is shortening at the end. Then came the cold winds which are separating autumn and winter. The fifth month which is May is the month that starts the winter season. There is no rain; the winds from the south are blowing strong. The trees shade their leaves; the grass becomes greyish, the wells become dry, the water freezes, the plants become dry at the farms. Then the Batswana person says the plnts are laughing at the birds and the birds do not eat those plants, now that is the May month.

Table 45: Concordance lines taken from the South African written Setswana corpus



MOTSHEGANONG: 38 entries (sort: 5L,5L)

N	Concordance	Set	Tag	Word No.	File	%
1	mo) 2. 3. Mopitlo 4. 5. Motsheganong 6. 7. Ph			1,120	us~1.txt	9
2	A re tsweletseng! a. Motsheganong yone, ke			1,058	walo2.txt	26
3	A re tsweletseng! a. Motsheganong yone, ke			2,931	walo2.txt	75
4	96, Port Elizabeth. 10 Motsheganong, 1965. K			23,152	tlha~1.txt	91
5	Port Elizabeth. 10 Motsheganong, 1965.			22,583	tlha~2.txt	90
6	e hard sorghum grains Motsheganong Motshe			31,411	cstt~1.txt	58
7	lepang ka kgwedi ya Motsheganong. E tlhaga			43,536	ojimst.txt	65
8	a dijalo go di segaka. Motsheganong o tlaa go			3,770	lofe~1.txt	43
9	go tloga ka kgwedi ya Motsheganong go fitlha			5,678	wemst.txt	15
10	nong Motsheganong Motsheganong motshe			31,413	cstt~1.txt	58
11	e di welang. Maloba fa Motsheganong e thulam			22,432	om~1.txt	48
12	ga mang? Botsenwa! Motsheganong a tihola			11,267	gwa~1.txt	96
13	joo. Mo kgweding ya Motsheganong Mafoko			9,493	dile~1.txt	48
14	otshediswamelelwane Motsheganong birds are			31,402	cstt~1.txt	58
15	edi o ntlele ka botshelo! Motsheganong ke ole o			1,985	oko~1.txt	38
16	a kgwedi, ka la di 31 Motsheganong, Khwela			14,125	tsw~2.txt	24
17	a kgwedi, ka la di 31 Motsheganong, Khwela			15,098	tsw~1.txt	27
18	phirima ka kgwedi ya Motsheganong. sele			41,912	anodi.txt	78
19	mabele ka kgwedi ya Motsheganong." Mah			23,711	swabil.txt	65
20	- .; nwe ke bonno. Motsheganong: Ke kgw			7,635	ana~1.txt	96
21	habula, Fa a omelela Motsheganong a fetoga.			4,277	tsa~1.txt	41
22	M~tseno a ~ago _ Motsheganong, Phuk			9,083	tsa~1.txt	88
23	gopolo e e phoso Mei Motsheganong meiyar			25,057	cstt~1.txt	46
24	mologong ya kgwedi ya Motsheganong ka Lama			18,961	wemst.txt	49
25	beilweng botsetse. Motsheganong kgwedi y			32,421	anodi.txt	60
26	-wa-godimo motshedi Motsheganong motshe			19,717	tcorel.txt	61
27	I. E ne e le kgwedi ya Motsheganong. Serame			45,429	ulemst.txt	90
28	ne ra 0 tlakaula mmidi Motsheganong - Ra s			6,736	\tsaya.txt	21
29	tlele ka kgwedi leina, Motsheganong e le leina			3,123	ana~1.txt	39
30	babetsa dinala serame Motsheganong: Mme			6,689	\tsaya.txt	21

When Tables 44 and 45 are compared, one notices the importance of gathering more information from the spoken corpus to add more value to the existing Setswana monolingual dictionaries.

5.6 Towards a sound lexicographical treatment of the Setswana months

Setswana months are named after nature and the change of the seasons. Every month falls within a season and it correlates with the historical events of that particular month as indicated above. According to the Batswana people, the first month of the year starts in August (*Phatwe)* and not January (*Ferikgong*) since they start with the plough and the first rain (*Kgogolammoko)* starts to fall. According to them, *Phukwi* (July) is regarded as the last month of the year and not December since all the work of the plough in the fields is completed. The meaning attached to these months is deduced from a specific cultural context. An illustration of that matter is given as follows:

Table 46: Treatment of the Setswana months

1	***Ferikgong***	*Go a fisa gore* ***logong lwa mofiri*** *lo robege bonolo. Ka jalo kgwedi ya nna Ferikgong, mm eke kgwedi ya bofelo ya setlha sa selemo.*
	January	**Eng.** The weather is very hot, drying up trees, especially those that produce firewood. It becomes easier to break firewood from such trees. It is the last month of Summer).
2.	***Tlhakole***	*Pula e na matsorotsoro, ka jalo e* ***tlhakola mogote*** *wa Ferikgong*

	February	**Eng.** During this month, very heavy rains are experienced, and those rains help to cool the hot January-weather off. It is the frost month of Autumn.
3.	***Mopitlwe***	*Dijalo le mmidi ke ntletsentletse. Ngwana wa motswana o tla ja mpa gore a* ***pipitlelwe,*** *ka jalo ya nna Mopitlwe.*
	March	**Eng.** During this month, it is green everywhere, the fields and the veld. There is also more than enough food for everyone. *(o tla ja mpa gore o pipitlelwe)*
4.	***Moranang***	*Go bonala go gola ga dijalo go anama. Go nama go iketla go namile maoto.* ***Moranang a dijalo a nama.***
	April	**Eng.** Plant growth slows down, some trees start losing leaves or their leaves start browning.
5.	***Motsheganong***	*Go tsididi, phefo e foka ka maatla, ditlhare di foforegi matlhare. Dinonyane ga di kgone go ka ja dijalo kwa masimong, ka jalo* ***Motswana a be a re dijalo di tshega nong*** *ka jalo kgwedi e bidiwa Motsheganong.*
	May	**Eng.** The weather starts getting cooler and windy, the trees/ plants starts losing leaves. Birds can no longer feed on the seeds easily in the fields, thus the Batswana mock (tshega) these birds frustration. It is the first month of winter.

6.	***Seetebosigo***	*Kgwedi e, **e kganela go eta bosigo** ka o tla tlhaela batho dikobo. Ka jalo **Seete bosigo**.*
	June	**Eng.** During this month, people are discouraged from visiting, especially for overnight/sleepover, because it is cold and guests might be an inconvenience if you do not have enough blankets to share with them (guests).
7.	***Phukwi***	*Ke kgwedi ya bofelo ya setlha sa mariga. Serame se a laela gore ke a feta. Go fitlha pula ya kgogolammoko. Motswana a bo a re Phukwa! A ngwaga o wele kwa! go tle o moša re tle re simolole botshelo seša. Go jewa dijo tse di bolelo le go apara diaparo tse di bokete.*
	July	**Eng.** The winter season comes to an end. It is also the end of the Batswana year. The first rains are expected in anticipation of the new year
8.	***Phatwe***	*Kgwedi e simolola setlha sa dikgakologo. Serame se a gakolologa. Ke kgwedi ya ntlha ya motswana.*
	August	Eng. This is the first month of the new year, and the first for the Spring season. The weather is slightly warmer and dusty but pleasant.
9.	***Lwetse***	*Ke kgwedi ya dikgakologo, botshelo jo boša bo a simolola, bo tla ka malwetse bo batho le diphologolo, ya nna kgwedi ya lwetse. Malwetse a mantsi a itemogelwa ka kgwedi e.*

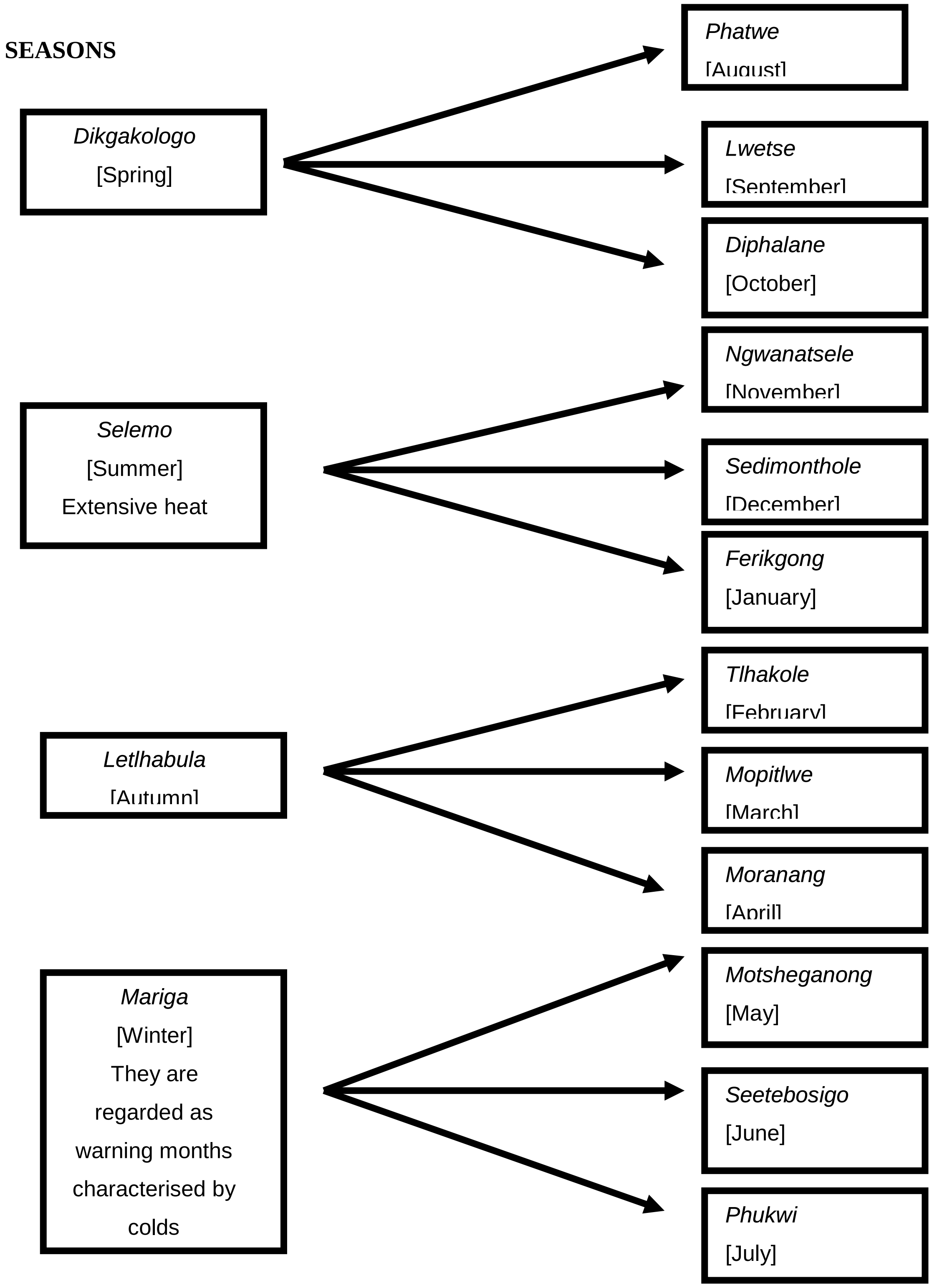
	September	**Eng.** This is a month for new beginnings, e, g. plants begin to blossom, most animals are now out of herbanation, etc. Also, because of the blossoming of plants, there are lots of pollen and allergens in the atmosphere, hence most diseases are spread and or experienced during this month.
10.	***Diphalane***	*Ke kgwedi ya dikgakologo. Bojang bo tlhogile, naga e talafetse, go utlwala melodi ya dononyane tse di itumeletseng bontle jwa naga ka bophara. Kgwedi eo e nne jaakadiphala mo tsebeng ya monna wa motswana.*
	October	**Eng**. This is the last month of Spring. It is green everywhere and birds are singing all over, the music is sweet to listen to.
11.	***Ngwanatsele***	*Ke tshimologo ya setlha sa selemo. Dophologolo do simolola go baya kgotsa go tsala. Kgwedi ya nna ngwanatsele. Bana ba itsela magapu le maugo. Ke ka moo go twang ngwana itseele.*
	November	**Eng.** This is the beginning of summer. Most animals give birth to their young ones, most fruit are ripe, and thus no one will stop you from having any fruit you want (ngwana itseele)
12.	***Sedimonthole***	*Maungo mo kgwedi e, a mantsi, merogo e simolola e jewa thata. Fa o ile sekgweng kgotsa masimong, o tla o rwele o imelwa ke dijo, o sa kgone le go 'ithola' morwalo. Ka jalo ra re 'sedi nthole ke palelwa ke go ithola' ka jalo kgwedi ya bidiwa*

		Sedimonthole.
	December	**Eng.** There are lots and lots of fruit and vegetables everywhere. For those who are reaping manually, using baskets would be so full, one won’t be able to remove the basket from one’s head by him. You might even wish that there was a bigger power to help you out.

According to the research, the names for Setswana months were shifted according to the historical events and the changes of seasons. It is important to state that a descriptive monolingual dictionary should be characterized by this kind of approach that will give true status of a language in question and THAN should not be an exception. This is about balance between the normal dictionary information and, where necessary, encyclopaedic information. The lexicographer should take a balanced approach in respect of the inclusion of encyclopaedic information in paper dictionaries. Due to the limitation on space encyclopaedic information should be limited to instances, e.g. crucial cultural information where the standard treatment in the article is insufficient. Paper Serswana dictionaries should therefore not be a combination of dictionary and encyclopaedia per se.

Table 47 below gives an overview representation of these Setswana months:

Table 47: Schematic representation of the Setswana months



SEASONS

Dikgakologo
[Spring]

Phatwe
[August]

Lwetse
[September]

Diphalane
[October]

Selemo
[Summer]
Extensive heat

Ngwanatsele
[November]

Sedimonthole
[December]

Ferikgong
[January]

Letlhabula
[Autumn]

Tlhakole
[February]

Mopitlwe
[March]

Moranang
[April]

Mariga
[Winter]
They are regarded as warning months characterised by colds

Motsheganong
[May]

Seetebosigo
[June]

Phukwi
[July]

5.7 Conclusion

In this chapter we have demonstrated the value of corpora, how it can be used to improve the quality of the microstructural elements in the treatment of lemma signs. The impact of corpora as a key to writing better dictionary articles, as sense distinctions for writing better definitions in monolingual dictionaries and setting up translation equivalents in bilingual dictionaries and as an aid to pinpoint frequent clusters was emphasized. We have also illustrated how a detailed analysis of corpus lines, in combination with frequency counts, enables lexicographers to tremendously enhance the quality of microstructural elements. Furthermore, typical collocations in addressing typical microstructural inconsistencies existing in the currently available Setswana dictionaries have been highlighted. Problems relating to the treatment of verbs, polysemy and the so-called synonyms in the Setswana dictionaries have been addressed. The study has also shown how the lexicographer with good query tools at his/her disposal can combine the output of different tools such as word-frequency counts and concordance lines screens.

We have also seen that the intuition of even a trained native-speaker lexicographer cannot compare to the accuracy of corpus-based queries. As a result corpus lexicography has the potential to result in much sounder and more user-friendly dictionaries than those compiled during the era of the so-called traditional manual lexicography. The treatment of the Setswana months in currently available Setswana dictionaries was critically analysed and evaluated against the background of the English and the Afrikaans dictionaries. The problem of what is regarded as the first month of the year for Setswana was outlined through the use of examples and diagrams. Suggestions for the improvement by means of a corpus-based microstructure has been identified and discussed in detail.

Chapter 6

Conclusion

We have entered the corpus era in the dictionary compilation and this study gave a comprehensive discussion of how future Setswana dictionaries should be compiled maximally using corpora, corpus query tools and advanced tools such as a ruler and block system. In chapter 2, the extensive historical background of the Batswana as a group (and how they are divided in both South Africa and Botswana), and Setswana as a language with diverse dialects was discussed in detail. It was also noted that Setswana, like any other language, is influenced by other languages spoken within or around its environment, unless where the environment is homogenously Setswana speaking where the influence is likely to be minimal or non-existent.

Just like other languages, Setswana is growing and this growth dates back to the arrival of the missionaries in the 1820's to date. The contribution made by the missionaries is valuable and worth noting. Various tables, maps, illustrations, etc, have been used to demonstrate the Batswana geographical locations, their language Setswana, dialects and their geographical distribution in both South Africa and Botswana, including the statistical representations of the Batswana and their dialects.

Language planning was also discussed in-depth, with emphasis on the three main sub-dimensions, characterizing language planning. Language is not stagnant, and thus language change also affects Setswana as a language. Factors like new development, technology, etc, are bound to affect language (particularly spoken language) and leads to the creation of new concepts and terms in relation to the new developments or for effective communication. New concepts usually originate from borrowing (from other languages) and analogical implications on some words frequently used. Just like many progressive languages, Setswana has grown and developed to be a language in its own right, through aspects such as the writing system orthography; which elaborated in-depth about the Setswana grammar, including the grammatical rules applicable. Various examples were given to illustrate and to support the discussions given.

The education system in South Africa, right from Bantu Education to date, greatly impacted on Setswana, particularly written or academic Setswana. Education systems around the world are prescriptive as to how language should be taught or learned – thus some words might be modified or disallowed completely, and some new (unfamiliar) words might be included. Language Boards also play a role in prescribing the parameters within which a language can operate, and some concepts might not be approved for use, as is the role of the PANSALB.

As the language develops, it becomes necessary to have some form of reference to keep up with the development, leading to the writing of dictionaries. In spite of the weakness of most of the Setswana monolingual dictionaries in comparison to English dictionaries, they (dictionaries) provide valuable information and serve as the basis for further language development. However, suggestions to deal with the weakness were given so that these dictionaries can be as useful as they were meant to be.

In chapter 3, the actual compilation of Setswana corpora was discussed in detail. Extremely useful theoretical insights of the strategies employed by COBUILD and LDOCE which are corpus-based are used as excellent examples for the compilation of African language dictionaries; in particular Setswana. Projects were used to illustrate the compilation of the Setswana oral corpus, and these beautifully illustrated the influence on Setswana language by various environmental aspects, such as the rural v/s the urban areas; level of literacy or education; attitudes to or against language changes; the origin of certain Setswana words (like Setswana months); and so forth.

The oral corpus serves as the foundation for the written Setswana, and the effect or role of spoken language has been highlighted effectively, covering various issues, e.g. keyness (both positive written corpus and negative corpus). The corpus query program was also used to give vital statistical information on the corpus. The statistical analysis between the oral corpus and the written corpus was highlighted. These analytical tests are applied in order to detect systematic differences between the text categories. Illustrations to indicate the statistical analyses were taken from WordSmith Tools, which is a very useful tool for distinguishing word types, tokens and the written corpus. The significance of frequency counts as an extremely useful tool in the compilation of a lemma list for a new dictionary was also emphasized.

The desire to redress the imbalance in favour of the unprejudiced investigation of the spoken and written language to justify its correctness was done by comparing the oral and written Setswana corpus in terms of the Keyness function using WordSmith Tools. Another very important comparison is given, between the South African corpus and the Botswana corpus. In this regard, words have been sorted by consistency, rank and overall frequency. However, the South African corpus is larger than the Botswana corpus as indicated through the illustrations used, and this difference is also evident with the statistical analysis. Words that are regarded as frequently used in the South African corpus are not necessarily regarded as frequently used in the Botswana corpus.

In chapter 4, much emphasis was placed on the urgent need for the utilization of electronic corpora on the macrostructural level. The macrostructural inconsistencies existing in Setswana dictionaries have been critically evaluated and analyzed. The use of frequency lists have been emphasized ensuring that frequently used words are not accidentally omitted and also that the dictionary space is not occupied by articles of lemmas unlikely to be looked for by the target users, particularly in the two Setswana monolingual dictionaries, i.e. THAN and THAND.

The limitations of the Setswana dictionaries are such that a huge number of words are entered without semantic information, thus the user has to frequently consult the front matter. This is a very time consuming and confusing process. Lexicographers therefore have to adopt a holistic approach when compiling dictionaries. This will also eliminate the possibilities of imbalances with regards to the alphabetical stretches, which are usually the result of the lexicographers adding new words as they come accross them, disregarding frequency counts. Sometimes lemma-signs are either under or over-treated. The reasons for this have been explored and graphical illustrations using the two well-known Setswana monolingual dictionaries were used.

The use of the multi-dimensional ruler to correct or balance the inconsistencies has been demonstrated. Lexicographers should negotiate a complex interplay and overlap between lemmatization strategies, approaches, lexicographic traditions, and verbal structures and conjunctiveness. Each of these were discussed in detail and typical examples provided or highlighted.

The rules for the lemmatization of nouns, both singular and plural, were looked into, once more referring to the two Setswana monolingual dictionaries. The problems experienced by users of these dictionaries or language users, in terms of the inconsistencies of some words converting to plural forms, have also been indicated and discussed to interpret the differences or inconsistencies for users familiar with the language and those that are not.

This chapter also opened new doors for the lemmatization strategies of nouns and verbs based on the frequency of use, using the available corpus data. The feasibility was exemplified through the macrostructural inconsistencies that exist in Setswana dictionaries. The problems inherent in the lemmatisation of the Setswana homonyms and tonal indication are real. These studies were performed against the background of the user- perspective.

Chapter 5 illustrates the contribution of corpora during the microstructural compilation of dictionary articles. Detailed analysis of corpus lines in combination with frequency counts at various levels was outlined through suitable examples taken from the Afrikaans and English dictionaries. The most significant corpus query output to generate the concordance lines and the occurrences of a word or phrase extracted from the corpus was emphasized. Such corpus lines will assist the lexicographer in respect of sense distinction, deciding on translation equivalents, retrieval of typical collocations, pinpointing frequent clusters and in the authentic examples to be included in a dictionary. Corpora in this instance can furthermore assist the lexicographer in finding typical collocations generated by WordSmith Tools and combination of words for writing a better definition or choosing better equivalents. Inconsistencies regarding the treatment of nouns, synonyms and cross-referencing as an interconnection of the microstructural component through the use of a reference marker were reviewed.

Bibliography

Dictionaries

(AED) The De Jager-Haum. 1980. Active *English Dictionary for English students*. London: Harrap Ltd.

(COBUILD) Sinclair, J.M. (Ed.) 1987. *Collins COBUILD English Language Dictionary*. London: Harper Collins Publishers.

(COBUILD) Sinclair, J.M. (Ed.) 1988. *Collins COBUILD English Language Dictionary*. London: Harper Collins Publishers.

(CNSD) Ziervogel, D. & Mokgokong, P.C. 1975. *Pukuntšu ye kgolo ya Sesotho sa Leboa, Sesotho sa Leboa-Seburu/Seisimane / Groot Noord-Sotho-woordeboek, Noord-Sotho-Afrikaans/Engels / Comprehensive Northern Sotho English-Afrikaans*. Pretoria: J.L. van Schaik.

(EZWB) Bryant, A.T. 1983. *English-Zulu Word-Book, Thirteenth edtion*. Mariannhill: Mission Press.

(KNSW) Ziervogel, D. & Mokgokong, P.C. 1988. *Klein Noord-Sotho woordeboek, N-Sotho-Afrikaans-English/Afrikaans-N-Sotho/English-N-Sotho*. Pretoria: J.L. van Schaik.

(MD) Eksteen, L.C. 1997. *Mayor Dictionary, Fourteenth Edition*. Cape Town: Boekhandel Ltd.

(MSED) Matumo, Z.I. 1993**.** *Setswana English Setswana English Dictionary*. Botswana: Macmillan Botswana.

(NSDN) Ziervogel, D. & Mokgokong, P.C.M. 1975. *Comprehensive Northern Sotho Dictionary, Northern Sotho-Afrikaans/English.* Pretoria: J.L. van Schaik.

(OALD5) Crowther, J. (Ed.) 1995. *Oxford Advanced Learner's Dictionary of current English*. Oxford: Oxford University Press

(OED) *Oxford English Dictionary, Second Edition*. 1989. Oxford: Oxford University.

(OSD) Hawkins, M. 1998. *Oxford School Dictionary*. Oxford: University Press.

(ONSD) De Schryver, G.M. (Ed.). 2007. *Oxford Bilingual School Dictionary*: *Northern Sotho and English.* (First edition.) Cape Town: OUP Southern Africa.

(PMEAD) Pharos. 2000. *Groot Woordeboek/ Major Dictionary. Afrikaans-English, English-Afrikaans*. Johannesburg: Pharos.

(PNSD) Kriel, T.J. 1988. *Popular Northern Sotho Dictionary*. Pretoria: J.L. van Schaik.

(PUKU 1) Kriel, T.J. 1983 *Pukuntšu Dictionary*. Pretoria: J.L. van Schaik.

(PUKU 2) Kriel, Theunis J. & Van Wyk, Egidius B. 1989. *Pukuntšu woordeboek, Noord-Sotho-Afrikaans, Afrikaans-Noord-Sotho*. Pretoria: J.L. van Schaik.

(SAOD) Mesthrie, R. 2004. *South African Oxford Dictionary*. Cape Town: Cape Town University.

(SEAD) Snyman, J.W. et al. 1990. *Dikišinare ya Setswana-English-Afrikaans Dictionary/ Woordeboek*. Pretoria: Via Africa.

(SED) Brown, J. 1985. *Secwana English Dictionary*. Botswana: Pula Press.

(SESD) Brown, J.T. 1987. *Setswana Dictionary*. *Setswana-English and English-Setswana*. Botswana: Pula Press.

(SETED) Brown, J. 1964. *Setswana-English Dictionary*. Harvard College Library Francisco: Wittenborn Art Books.

(SZD) Dent, G.R. & Nyembezi C.L.S. 1993. *Scholar's Zulu Dictionary*. Pietermaritzburg: Schuter & Shooter (Pty) Ltd.

(THAN) Kgasa, M.L.A., & Tsonope, J. 1995. *Thanodi ya Setswana*. Botswana: Longman.

(THAND) Kgasa, M.L.A. 1976. *Thanodi ya Setswana ya Dikole*. Cape Town, South Africa: Longman Penguin South Africa (Pty) Ltd.

(TNODE) *The new Oxford dictionary of English*. 1998. New York: Oxford University Press.

(TSAOSD) Hawkins, J.M. 1996. *The South African Oxford School Dictionary*. Oxford: University Press.

(VAW) Labuschagne, F.J. & Eksteen L.C. (Ed). 1993. *Verklarende Afrikaanse Woordeboek*. Pretoria: J.L. van Schaik.

Other literature

AL-Kasimi, A.M. 1983. *Linguistics and Bilingual Dictionaries*. Leiden: E.J. Brill.

Allen, R. 2006. *Grading the corpora for language learning*: the effect lexical chunks paperm read at the Third international IVACS conference, 24 June 2006.
Available at http:// www. Nottingham.ac.uk/English/IVACS/allan.pp.

Andor, J. 1989. *Strategies, tactics and realistic methods of text analysis*. In W. Heydrick, F. Neubauer, J. Petofi, and E. Sëzer (Eds.), Connexity and Coherence Analysis of text and discourse (pp. 28-36). Berlin: Walter de Gruyter.

Artevelde.T. AFRILEX-REEKS 12:2002. *African Association for Lexicography*. Stellenbosch Harteveld T. Afrilex-Reek 11:2001 African Association for Lexicography. Buro van die Wat Stellenbosch.

Atchison, J. 1981**.** *Language Change*. *Progress or Decay*? London: Hutchison.

Atkins, B.T. 1985. *Monolingual and Bilingual Learners' dictionaries.* A Comparison.

Atkins, B.T. Sue, (Ed.) 1997. *Bilingual dictionary data type*. Oxford: Oxford Press.

Atkins, B.T. Sue, (Ed.) 1997. *Monolingual dictionary data type*. Oxford: Oxford Press.

Atkins, B.T. Sue, (Ed.) 1988. *Using Dictionaries. Studies of Dictionary use by Language Learners and Translators.* Tubingen: Max Niemeyer Verlag.

Barber, C.L. 1972. *The story of language, revised*. London and Sydney: Pan Books.

Barlow, M. 2000. *MonoConc Pro Concordance software*. Houston: Athelstan.
Available at www. Michaelbarlow. Com/viz. Html.

Baroni, M. & Bernadini, S. 2004. Bootstrapping corpora and terms the Web. In: *Proceeding of the 4th International Conference on language Research and evaluation.* LREC-2004, Lisbon.

Biber, D. 1993. *Using register-diversified corpora for general language studies.* Association for computational Linguistics 19(2): 219-242.

Biber, D. Conrad S. & Reppe R. 1998. *Corpus Linguistics Investigating Structure and Use*. Cambridge: Cambridge University.

Breutz, P.L. 1989. *History of the Batswana*. Natal: Ramsgate, R.S.A.

Brown, R. & Lunenburg, E.H. 1954. *A study of language and cognition, Journal of Abnormal and Social Psychology 49:452-60*. Reprinted in Brown: Saporter.

Brumfit, C. 1992. Language Planning. *The Oxford Companion to the English Language*. Ed. Tom McArthur. Oxford: Oxford University Press.

Busane, M. 1990. *Lexicography in Central Africa; the user perspective with special reference to Egypt*. Hartmann, R.R.K. (Ed). In Lexicography in Africa. Volume 15, Exeter Linguistic Studies: University of Exeter Press.

Carroll, J.B. 1953. *The study of language.* Cambridge: Cambridge University Press.

Chaucer, G. 1902. *The Cambridge History of English and American Literature*. United States: Cambridge University.

Ciaramita, M. and Baroni, M. 2006. *Measuring Web-Corpus Randomness: A progress Report*. Italy: Universita di Bologna.

Cole, P., and Morgan, J.L. (Eds). 1975. *Syntax and Semantics, Vol.3*: Speech Acts. New York: Academy Press.

Cole, D.T. 1992. *An introduction to Tswana Grammar*. Cape Town and Johannesburg: Longman.

Cole, D.T. 1955. *An introduction to Tswana Grammar, 9th impression*. Cape Town: Longman.

Cooper, R.L. 1989. *Language planning and* social *change*. New York: Cambridge University Press. Avilable at http://en.wikipedia.org/wiki/Language_planning
Retrieved: 2008-06-20.

Crawford, J. (in press). *Endangered Native American languages*. What is to done, and why? In Recento and Burnaby (Ed.). Language and policies in the US and Canada: Myths and realities. Mahwah, NJ: Lawrence Erlbarum Associates.

Crawford, J. 1995. *Seven hypotheses on language loss*: Causes and cures. Paper adapted from a speech at the second symposium on stabilizing indigenous languages. Northern Arizona University, Flagstaff, AZ.

Crick, M. 1976. Exploration in language and meaning. *Towards a semantic Anthropology*. London: Malaby.

Crystal, D. 1985. A dictionary of linguistics and phonetics. 2nd edition. New York: Basil Blackwell. Available at www. Sil.org/Linguistics/Bibliography Linguistics/ Crystal1985.htm

Crystal, D. 1986. *The Ideal Dictionary, Lexicographer and User*. Ilson, R. (Ed.) Lexicography: An emerging International Profession: 72-81. Manchester University Press.

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

De Schryver, G.M. & Prinsloo, D.J. 2000b. Electronic corpora as a basis for the compilation of African-language dictionaries, part 2: The microstructure. *South African Journal of African Languages* 20/4: 310 – 330.

De Schryver, G.M. & Prinsloo, D.J. 2001. Corpus-based Activities versus Intuition-based Compilations by Lexicographers, the Sepedi Lemma-Sign List as a Case in Point. *Nordic Journal of African Studies 10/3: 374-398.*

Diamond, J. 1993. *Speaking with single tongue*. Discover pp. 78-85.

Doke, C.M. & Cole, D.T. 1961. *The History of Bantu Linguistics.* Johannesburg: Witwatersrand University Press.

Fox, G. 1987. The Case for Examples. In Sinclair, J. (Ed). *Looking up: An account of the COBUILD project in lexical computing*. London: Collins ELT.

Galley, M. 2000. *The Journal of Computer-Mediated Communication*. Washington: Department of communication.

Gauton, R. & Taljard, E. Honours course: *Comparative Bantu Linguistics*. Pretoria: University of Pretoria.

Gouws, R.H. 1983. *Problems of Neologisms in Nothern Sotho, Information Categories in Dictionaries with Special Reference to South Africa*. Pretoria: J.L. Van Schaik.

Gouws, R.H. 1990a. *“Information categories in dictionaries, with special references to Southern Africa”*. Hartmann, R.R.K. (Ed.). Lexicography in Africa, Volume 15, Exeter Linguistic Studies. Cambridge: University of Exeter Press.

Gouws, R.H. 2000b. Towards the Formulation of a Metalexicographic. *Founded Model for National Lexicography Units in South Africa.* In Wiegand, Herbert E. ed. 2000. Wörterbücher in Der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium (Lexicographical Series Mayor 100): 109-33. Tubingen: Max Niemeyer Verlag.

Gouws, R.H. & Prinsloo, D.J. 1998**. "***Cross-referencing as a Lexicographic Device."* Lexicons 8: 17-36.

Gouws, R.H. & Prinsloo, D.J. 2005. '*Left- expanded Article Structures in Bantu with Special Reference to Isizulu and Sepedi*.' *International Journal of Lexicography* 18: 25-46.

Gumperz, J. (1982). *Wikipedia, the free encyclopaedia USA*: University of California. Available at http://en.Wkipedia.org/wiki/John.J.Gumperz.htm

Hartmann, R.R.K. 1989. *Lexicographers and their work. Volume14*. Cambridge: University of Exeter Press.

Hartmann, R.R.K. 1990a. *Lexicography in Africa*. *Volume 15*. Cambridge: University of Exeter Press.

Hartmann, R.R.K. 1990b. *Dictionary workbooks. Volume 16*. Cambridge: University of Exeter Press.

Haugen, E. 1966. *Dialect, Language, Nation*. Cambridge: Harvard University Press.

Hausmann, F.J., Reichmann, O., Wiegand, H.E. and Zgusta, L. (Eds.). *An International Encyclopaedia of lexicography Dictionaries*. New York: Walter de Gruyter. Available at http://en.Wikipedia.org/wiki/Language development.htm

Hoey, M. 1991. *Patterns of Lexis in the text. Oxford*: OUP.

Huebner, T. 1987. *A socio-historical approach to literacy development*. A comparative case study from the Pacific. In J.A. Langer (Ed.). Language, literacy, and culture: Issues of society and schooling (pp. 178-196). Norwood, New Jersey: Ablex Publishing Co.

Hymes, D.H. (Ed.) 1964**.** *Language in culture and Society*. New York: Harper & Row.

Ilson, R. (Ed.) *Dictionaries, Lexicography and Language Learning. 15*-24. Oxford: Pergamon Press.

James, Gregory, Davison, Roberts, Cheung Heung-jeung, Amos and Deerweester, Scott. 1994. *English In Computer Science, Acorpus-Based Lexical Analysis*. Hong Kong: Longman Asia Limited.

Johansson, S. 1984. *Looking up, Frequency analysis of English Vocabulary and grammar*. Clarendon: Press Oxford.

Johansson, S. and Hofland, K. 1989. *Frequency analysis of English vocabulary and grammar based on the LOB corpus*. Oxford: Oxford University Press.

Kennedy, G. 1998a**.** *An introduction to corpus Linguistics*. New York: Addison Wesley Longman.

Kennedy, G. 1998b. *An introduction to Corpus Linguistics*. Paris: Limsi-NRS and Universite' Paris.

Kilgarriff, A. 1977a. Putting *Frequencies in the Dictionary*. International Journal of Lexicography 10(2): 135-155.

Kilgarriff, A. 2001. Comparing corpora. *International Journal of corpus Linguistics* 6 (1, 1-37).

Kilgarriff, A., and Grefenstette, G. 2003. *Introduction to the Special Issue on the Web as Corpus*. Computational Linguistics 29(3): 323-347.

Kintsch, W. and Van Dijk, T. 1978. *Towards a Model of Text Comprehension and Production keyness*. Cambridge: Cambridge ESOL.

Kriel, T.J. 1976. *The new English-Northern Sotho, Northern Sotho-English*. Johannesburg: Educum Publishers.

Kruger, J.H. 1965. *Afrikaans loanwords in Setswana languages*. Cape Town: Oxford University.

Kucera, H. and Francis W.N. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Kulick, D. 1994. *Language shift and cultural change*. Paper presented at La Trobe University: Australian Linguistic Institute on language maintenance and shift, Victoria, Australia

Lancashire, P.D. 1993. *Cognitive factors in source monitoring and auditory hallucinations*. London: Cambridge University Press.

Landau, S.I. 2001. *Dictionaries*: *The Art and Craft of Lexicography (2nd edition).* Cambridge: Cambridge University.

Langacker, R.W. 1987. *Foundation of cognitive Grammar* (Volume 1.) Stanford: Stanford University Press.

Laufer, B. 1992. *Corpus-based versus Lexicographer Examples in Comprehension and Production of New* Words. Tommola, Harnu et al. Cambridge: Cambridge University Press.

Leech, G. 1981. *Semantics***.** England: Clays Ltd

Leech, G. 1990. *Semantics*. England: Penguin Book Ltd.

Leepile, T. 2003. *Languge and National Development in Proceeding of the congress of the Language*. Botswana: The Business of Newspaper.

Lyons, J. 1981. *Language, Meaning and Context*. London: Fontana / Collins.

Lyons, J. 1990. *Language and Linguistics*. Cambridge: Cambridge University Press.

Lyons, J. 1992. *Introduction to theoretical Linguistics*. Cambridge: Cambridge University Press.

Mansour, G. 1993. *Multilingualism and national building*. Clevedon, England: Multilingual Matters Ltd.

Martin et al. 1983. *On the Processing of a Text Corpus, From textual data to lexicographical information*. In Reinhard R.K. Hartmann (ed.), pp 77-87.

Mills, J. 1998. *Lexicon Based Critical Tokenisation*: An Euralex'98, pages 213-220, August.

Moloto, D.P. No date. *Motimedi*. Johannesburg: Bona Press.

monoConc www. Michaelbarlow. Com/viz.html and www.lexically. net/ download/ version4/htm/index.htm

Moon, R. 1998. Fixed Expression and Idioms in English: *A corpus-based Approach*. Oxford: Oxford University.

Prinsloo, D.J. 1994. *Lemmatization of verbs in Northern Sotho*. S. A. Journal of African Languages 14(2): 93-102.

Prinsloo, D.J. & De Schryver, G.M. 1999**.** *The lemmatization of nouns in African Languages with special reference to Sepedi and Cilubà***,** *South African Journal of African Languages*, 19(4): 258-75.

Prinsloo, D.J. and De Schryver, G-M. (eds.) 2000. SeDiPro 1.0, *First Parallel Dictionary Sepedi-English*. Pretoria: University of Pretoria.

Prinsloo, D.J. & De Schryver, G.M. 2001a. *Monitoring the stability of a Growing Organic Corpus, with special reference to Sepedi and Xitsonga Dictionaries*: *Journal of the Dictionary Society of North America* 22:85-129, 2001.

Prinsloo, D.J. & De Schryver, G.M. 2001b. *Corpus applications for the African languages, with special reference to research, teaching, learning and software.* South African Linguistics and Applied Language Studies 19/1-2; 111-131.

Prinsloo, D.J. & De Schryver, G-M. 2003. Effective vordering met die *Woordeboek van die Afrikaanse Taal soos gemeet in term van n multidimensionele Liniaal* [Effective Progress with the *Woordeboek van die Afrikaanse Taal as Measured in terms of a Multidimensional Ruler*]. Botha, W.F. (Ed.). 2003. ‘n Man beur. Huldigingsbundel vir Dirk van Schalkwyk 106-126. Stellenbosch: Buro van die WAT.

Prinsloo, D.J. 2004. *Revising Matumo's* Setswana-*English-Setswana Dictionary*. Lexikos, 14 (158-172). Departments of African Languages, University of Pretoria.

Prinsloo, D.J. & Gouws, R.H. 1995. *Principles and Practice of South African Lexicography*. Stellenbosch: African Sun MeDIA.

Prinsloo, D.J. & Gouws, R.H. 1996. *Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho, South African Journal of African Languages,* and 16(3): 100-7.

Prinsloo, D.J. & Gouws, R.H. 2000. *The use of examples in polyfunctional dictionaries*. Lexikos, 10. pp 138-156.

Prinsloo, D.J. & Gouws, R.H. 2005. *Lemmatization of Nouns in African languages wuth special reference to Sepedi and Ciluba*. Pretoria: Department of African Languages, University of Pretoria.

Prinsloo, D.J. 2009. *Traditions, trends and changes in lexicography*. Pretoria: University of Pretoria.

Prinsloo, D.J. & Heid, U. Forthcoming: *A bilingual dictionary for a specific user group: supporting Setswana speakers in the production and reception of English*. Pretoria: Department of African Languages, University of Pretoria.

Ramagoshi, R.M. 2000. *History of Setswana as a written language*. Pretoria: Pretoria University.

Republic of Botswana. 2001. Central statistics Office 2001 Census. Available at http://www.cso.gov.bw Retrieved: 2007-02-15.

Renouf, A. 1987. Corpus Development. In Sinclair. John M. (Ed). 1987**.** *Looking up, an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary* 1-40. London: Collins ELT.

Rundell, M. 1996. *The corpus of the future and the future of the corpus*. Talk at iniversity of Exeter special conference, [cited 17 July 2004. Available from http// www. Ruf..rice. edu/ ~ Barlow/ future. html.

Schapera, I. 1953. *Bantu-speaking Tribes of South Africa*. London: Routledge and Kegan Paul Ltd.

Scott, M. & Tribble, C. 1996. Textual Patterns. *Key words and corpus analysis in language education*. Amsterdam/ Philadelphia: University of Liverpool and King's College, London University.

Scott, M. & Tribble, C. 1999. *WordSmith Tools Users Help File*. Oxford: OUP.

Shannon, C.E. 1948. *Literature on the issue of representativeness of corpora*. Oxford: Oxford University Press.

Sinclair, J.M. (Ed.) 1987. *Looking up, An account of the COBUILD Project in lexical Computing and the development of the COBUILD English* Language *Dictionary*, London: Collins ELT.

Sinclair, J.M. 1991. *Corpus, concordance, Collocation*. London: Oxford University Press.

Spencer, M. (in press). And what of the language of Micronesia. In Mugler, F. and Lynch, J. (Ed.). *Pacific languages in Education*. (Pp. 10-35), Suva, Fiji: Institute of Pacific Studies.

Summers, D. 1993. *Longman/ Lancaster English Language Corpus- Criteria and Design*, International Journal of lexicography 6(3): 181-208.

Tauli, V. 1964. Practical Linguistics: The Theory of Language Planning. *Proceedings of the Ninth international Congress of Linguists*. Ed. Horace, G. Lunt. The Hague: Mouton & Co: 605-609.

Tauli, V. 1968. *Introduction to a theory of language planning*. Uppsala: Almqvist and Wiksells. Available at http:// en. Wikipedia. Org/ wiki/ Language planning.htm Retrieved: 2007-02-16.

Tsonope, J. 1990. *Birth of the National Language- The history of Setswana Language*. Pretoria: Via Africa.

Tswana Terminology and Orthography No. 3. South Africa: Department of Education and Training.

Tswana Terminology and Orthography No. 4. South Africa: Department of Education and Training.

Tshetlo, S.S. 1984. *Moratho o montsho*. Babelegi: Craft Press.

UNESCO. 2000. *World Language Report Survey Questionnaire*. Pretoria: National Language Service Department of Arts, Culture, Science and Technology. Available at http//www.cyberserv.co.za/users/~jako/lang/unesco/Setswana.htm. Retrieved: 2007-11-01.

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

Williams, R. 1983. Keywords. *A Vocabulary of Culture and Society*. London: Fontana.

Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

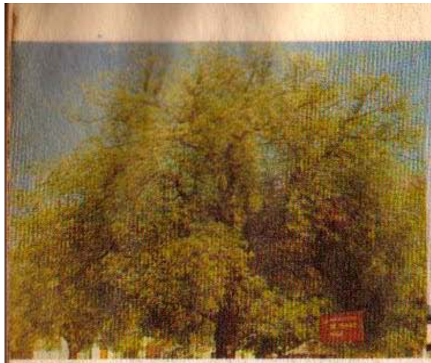
WordSmith Tools www.lexically.net/ wordsmith/corpus _ linguistics _ links/Wilkinson.doc.

Zgusta, L. 1971. *Manual of Lexicography*. The Hague: Mouton.

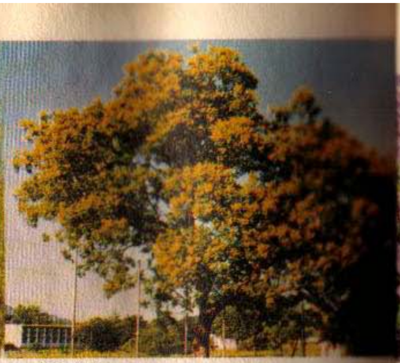
APPENDIX 1: Summary of the 13 oral Setswana projects

		Tokens	Types
Project	1	1,740	416
Project	2	2,526	598
Project	3	3,778	1,176
Project	4	1,520	419
Project	5	3,197	521
Project	6	3,414	734
Project	7	4,384	1,203
Project	8	827	354
Project	9	4,736	1,024
Project	10	2,754	716
Project	11	1,190	327
Project	12	1,940	554
Project	13	1,217	395

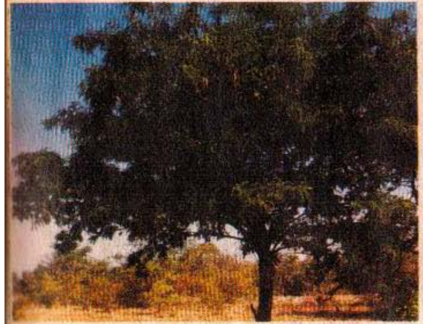
APPENDIX 2: Pictures of trees in Thanodi ya Setswana



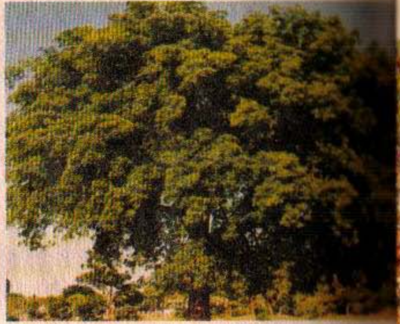
Mokala



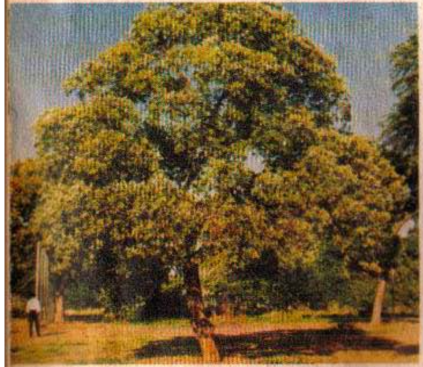
Mosêtlha



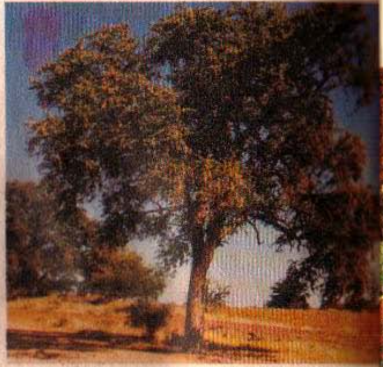
Modubana



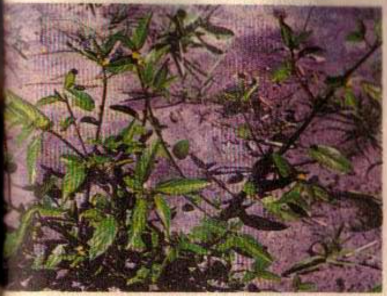
Morula



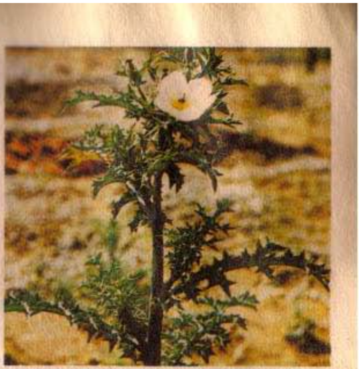
Mogônônô



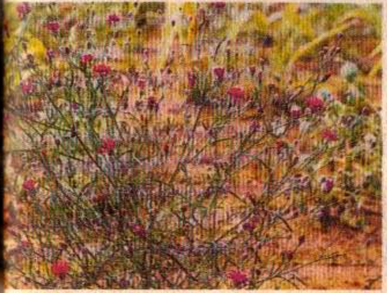
Motswere



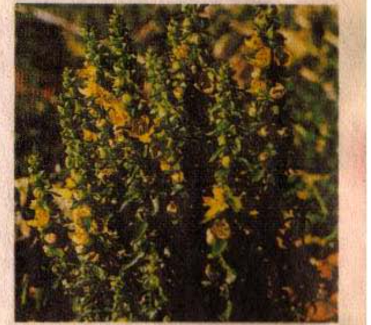
lele



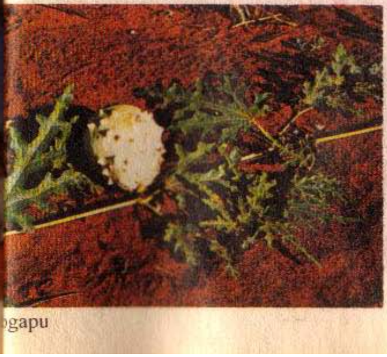
Lopêro



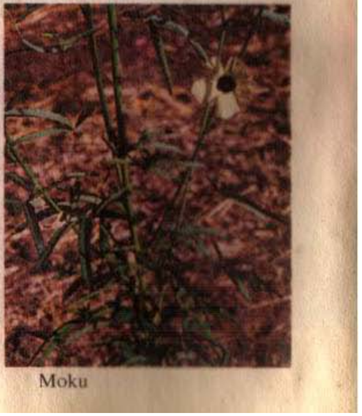
goleng



Matebele



gapu



Moku