UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Corpus design for Setswana lexicography

Thapelo Joseph Otlogetswe

A thesis submitted in accordance with the requirements
for the degree of Ph.D. in African Languages at
The University of Pretoria.
September, 2007.

Promoter: **Prof. D.J. Prinsloo**
Co-Promoter: **Dr. Adam Kilgarriff**

Summary

This PhD thesis is about the design of a Setswana corpus for lexicography. While various corpora have been compiled and a variety of corpora-based researches attempted in African languages, no effort has been made towards corpus design. Additionally, although extensive analysis of the Setswana language has been done by missionaries, grammarians and linguists since the 1800s, none of such research is in corpus design. Most research has been largely on the grammatical study of the language.

The recent corpora research in African languages in general has been on the use of corpora for the compilation of dictionaries and little of it is in corpus design. Pioneers of this kind of corpora research in African languages are Prinsloo and De Schryver (1999), De Schryver and Prisloo (2000 and 2001) and Gouws and Prisloo (2005).

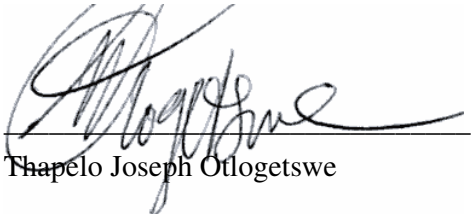
Because of a lack of research in corpora design particularly in African languages, this thesis is an attempt at filling that gap, especially for Setswana. It is hoped that the finding of this study will inspire similar designs in other languages comparable to Setswana.

We explore corpus design by focusing on measuring a variety of text types for lexical richness at comparable token points.

The study explores the question of whether a corpus compiled for lexicography must comprise a variety of texts drawn from different text types or whether the quality of retrieved information for lexicographic purposes from a corpus comprising diverse text varieties could be equally extracted from a corpus with a single text type. This study therefore determines whether linguistic variability is crucial in corpus design for lexicography.

Declaration

I declare that **Corpus design for Setswana lexicography** is to the best of my knowledge and belief, my original work. All the sources that I have used or quoted have been indicated and acknowledged by means of complete references. The material has not been submitted, either in whole or part, for a degree at this or any other university.



Thapelo Joseph Otlogetswe

Acknowledgements

Heartfelt gratitude to my supervisors:

- **Dr. Adam Kilgarriff.** I first met Dr Kilgarriff's research as a postgraduate student at the University of Oxford. Since then I wanted to be his student. It has been a wonderful and enriching experience to study under his excellent supervision.

- **Dr. Roger Evans.** Before my transfer from the University of Brighton (UK) to the University of Pretoria (SA), Roger Evans was one of my supervisors. I am exceedingly grateful for his guidance in the earlier stages of this study.

- **Prof. Daan Prinsloo.** I am grateful to Prof Prinsloo's exceptional supervisory leadership and patience with editing my work and for access to Setswana corpora.

.... Many thanks also to colleagues at ITRI University of Brighton, particularly Ying Ling. ITRI provided me with the finest research atmosphere for my study to flourish earlier in my PhD.

… Many thanks to Steve Crowdy (Longman dictionaries, UK) who shared the documentation of the BNC spoken corpus design and transcription with me.

… Many thanks to Mike Scott who availed many statistical papers and clarification of some of the programming behind Wordsmith Tools 4.

.... I am indebted also to my sponsor, The University of Botswana for funding the larger part of my PhD.

.... I am equally grateful to the committee Overseas Research Students Awards Scheme (ORS) administered by Universities UK committee for selecting me for one

of their awards. The award paid for part of my tuition at the University of Brighton. The award was given on a competitive basis to international postgraduate research students of outstanding merit and research potential.

.... Thanks to the following for availing the Setswana text for inclusion in the corpus

- Prof. D.J. Prinsloo, University of Pretoria.
- Botswana Macmillan.
- Botswana Parliament.
- *Mokgosi* newspaper.
- Many secondary schools whose identity we are not disclosing in the interest of anonymity.
- Department of Information and Broadcasting.
- Different Botswana government departments.
- Prof. Kevin Patrick Scannell of the Department of Mathematics and Computer Science, Saint Louis University for helping with harvesting Setswana text on the Web.
- Dr Elma Thekiso (University of Botswana) who was kind enough to give us her court transcriptions.
- Different families and individuals who allowed us to tape their conversations.
- Thanks to Motlhaleemang Ntebela for giving us text from the *Mmegi* newspaper.

.... I am thankful to my wife, Shinie Otlogetswe who has been a great source of support during some of the most difficult times of this study.

.... Finally, my unwavering faith in God, through Jesus Christ, has remained a rock and encouragement throughout this study.

Abbreviations

BNC	British National Corpus
CDIF	Corpus Development Interchange Format
CI	Confidence Interval
CLAWS	Constituent Likelihood Automatic Word-tagging System
COBUILD	Collins Birmingham University International Language Database
CQS	corpus querying software
DDP	Dictionary Development Process
HLT	Human Language Technology
HTML	Hyper-text mark-up language
JFIT	Joint Framework for Information Technology
KWIC	Key Word in Context
LDOCE	Longman Dictionary of Contemporary English
LMS	London Missionary Society
LOB	Lancaster/Oslo-Bergen Corpus
MD	Multi-Dimensional
MWE	multi-word expression
NLP	Natural Language Processing
OED	Oxford English Dictionary
POS	Part of speech
RNPE	Revised National Policy on Education
RRC	Russian Reference Corpus
SCA	Simple Consistency Analysis
SDA	Seventh Day Adventist
SGML	Standard Generalized Mark-up Language
SIL	Summer Institute of Linguistics
STTR	Standardized type/token ratio
TEI	Text Encoding Initiative
TSB	Traditional Setswana Beliefs
TTR	Type/token ratio
WWW	World Wide Web
XML	Extensible Mark-up Language

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Background to the study

This thesis is about corpus linguistics, precisely corpus design for lexicography (the science and art of dictionary compilation) as it relates to the Setswana language. The field of corpus linguistics is broad, covering areas such as grammatical studies, language education sociolinguistics, phonetics, phonology, stylistic analysis, dialectology and others (Kennedy, 1998). Corpus linguistics, particularly its application to lexicography is in its infancy in many African languages, particularly so in the language which is the focus of this thesis: the Setswana language. The larger body of Setswana research and that of many African languages covers broad linguistic areas such as language attitudes and use (Savage, 1990; Mooko, 2002; Bagwasi, 2003), language ecology (Anderson and Janson, 1997), grammar (Cole, 1955), syntax (Demuth and Johnson, 1989) phonology and phonetics (Jones and Plaatjie, 1916/1928; Mathangwane, 2002; Chebanne, 2002), and language literacy (Molosiwa, 2004).

Almost all of the studies mentioned in the preceding paragraph do not use corpora. Those that use corpus data are in the minority and relate to the use of corpora for lexicography. Amongst these are Prinsloo and Gouws (1995) Gouws and Prinsloo (1997) Prinsloo and De Schryver (1999) and Prinsloo (2004). Furthermore most research in corpora for the African languages is aimed at the compilation of corpora for lexicographic use and not in corpus design. This study focuses on Setswana corpus design whose output can serve a lexicographic purpose. Its findings and methodologies it is hoped would inspire similar designs in other African languages.

In corpus research in general, the focus has been placed on what researchers can

retrieve from corpora, amongst these being frequency information, lemma lists, example sentences in dictionaries and concordance lines (De Schryver, 2002: 275/6). While there is nothing defective with such studies, what is lacking in the literature is detailed and in depth research on corpus design particularly for African languages. The gap is particularly worrying in that the quality of corpus output is dependant on corpus design.

Few corpus designs have been documented. Francis and Kucera (1982) document the meticulous nature of the Brown Corpus design, while Crowdy (1991, 1993 and 1994) discusses in detail the sophistication of the British National Corpus spoken component compilation and Burnard (1995) outlines the design of the entire British National Corpus. On the basis of what has gone into such corpora, researchers are able to determine how valuable corpus output of such corpora is. In our research we have not found any study in corpus design which outlines the design of any corpus in African languages. This thesis' objective, as will be outlined below, in part is to fill this gap.

1.2 Statement of the research problem

Corpora use is not common in many dictionary projects in Africa languages, Setswana included. The larger body of research in corpora is on corpus usage and rarely in corpus design. There is no research that focuses on the design of Setswana language corpora.

At a practical lexicographic level, the production of dictionaries in various African languages has been very low particularly when compared with dictionary compilation in English by publishing houses such as Oxford University Press, Longman, Webster, COBUILD (The Collins Birmingham University International Language Database) and Chambers. For instance since 1875 less than ten Setswana dictionaries have been compiled. Three of these are monolingual dictionaries (Kgasa, 1976; Kgasa and Tsonope, 1998 and Dent, 1992), one is trilingual (Snyman et al., 1990), and three are bilingual (Brown, 1925, Matumo, 1993 and Créissels and Chebanne, 2000). More dictionaries could have been compiled considering that Setswana has official status in

South Africa and it is Botswana's national language (and not its official language as Onibere et al. (2001: 503) claim). None of the Setswana dictionaries mentioned above used corpora save for Kgasa and Tsonope (1998).

At a theoretical level, several corpus design issues are still to be explored. The question of how corpora should be compiled as resource bases for lexicography is still to be sufficiently researched. There is therefore a need to measure how best to design corpora whose output will closely reflect the character of the varieties of Setswana as they are used. At the centre of this thesis, therefore, is the question: what kind of corpus is 'better suited' for Setswana lexicography? The question translates into the following issues:

1. Which text types exist in the Setswana language? In which contexts is the language used? These questions are significant since what we wish to establish is the language text types that could be added to the compilation of a corpus. Beyond that, experimentally we want to calculate and measure which words are typical of a text type.

2. The lack of structured corpora on which experiments can be conducted remains a huge problem for many languages. In many cases of African languages there are no corpora, and in cases where they exist, they are usually purely *opportunistic*; a simple gathering of whatever text exists without an attempt of representing language variability in the structure of the corpus. The question that needs addressing is therefore, how best to compile a corpus or corpora for Setswana lexicography but also for other Human Language Technology (HLT) and Natural Language Processing (NLP) purposes which capture the linguistic variability of the language. Additionally, what types of language components should go into the corpus composition and in what quantities? Finally, how can we empirically account for what constitutes corpora for lexicography in Setswana?

1.3 Clarifying terms: genre, text type and varieties

Before we proceed further in this thesis, it is important that we briefly define the terms: text types, genre and varieties which are sometimes used differently in the literature. We discuss how various scholars use the terms and how the terms are used in this study. Genre has been defined thus:

> …texts that have a similar set of purposes, mode of transmission and discourse properties (Roberts, 1998: 79).

> …a category assigned on the basis of **external** criteria such as intended audience, purpose, and activity type, that is, it refers to conventional, culturally recognised groupings of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the **internal** (linguistic) criteria forming the basis of text type categories (Lee, 2001: 38; emphasis in the original).

> Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic; they are assigned on the basis of use rather than on the basis of form (Biber, 1988: 170)

Bussmann defines text types

> ….a term from **text linguistics** for different classes of **texts**. Within the framework of a hierarchical text typology, text types are usually the most strongly specified class of texts (e.g. recipes, sermons, interviews), characterised by different internal and external features (Bussmann, 1996: 481/2).

He also defined linguistic variety as,

> … a generic term for a particular coherent form of language in which specific extralinguistic criteria can be used to define it as a variety. For example, a

> geographically defined variety is known as a **dialect**, a variety with a social basis as a **sociolect**, a functional variety as a jargon or a **sublanguage**, a situative variety as a **register** (Bussmann, 1996: 512).

> One way of making a distinction between *genre* and *text type* is to say that the former is based on external, non-linguistic, "traditional" criteria while the latter is based on the internal, linguistic characteristics of texts themselves. A *genre*, in this view, is defined as a category assigned on the basis of external criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the internal (linguistic) criteria forming the basis of *text type* categories (Lee, 2001: 38).

Lee also argues that genre and register overlap:

> The two terms *genre* and *register* are the most confusing, and are often used interchangeably, mainly because they overlap to some degree. One difference between the two is that *genre* tends to be associated more with the organisation of culture and social purposes around language and is tied more closely to considerations of ideology and power, whereas *register* is associated with the organisation of situation or immediate context. Some of the most elaborated ideas about genre and *register* can be found within the tradition of systemic functional grammar ((Lee, 2001: 41/42).

Some linguists make distinctions between genres, domains and text types, as in Lee (2001). In this thesis such distinctions are not applied, instead we use genre, text types and varieties inter-changeably to refer to linguistic variability in general. Our position is similar to that of Aston (2001: 73) who uses the term “the term "text type" as a neutral one which does not imply any specific theoretical stance” but rather in general to refer to linguistic variability.

1.4 Methodology

There is a large body of lexical research which deals with comparing different language varieties to measure language variation or describe lexical qualities of a subcorpus (Biber, 1993; Kilgarriff, 1996, 1997a; Leech et al., 2001; Sharoff, 2006). Other comparisons and measurements have been done at the level of corpora (Kilgarriff and Salkie, 1996) where corpora have been compared for similarity and homogeneity through word frequencies. To achieve such comparisons for lexicography there is a need for large corpora that cover substantial samples of each significant variety of a language, so that the lexicographer does not miss words or patterns of word use from a variety of genres (Biber, 1990: 263).

However, what such varieties are and in what proportion they have to appear in a corpus is usually not clear. Central to our argument in this thesis is that the capturing of different varieties in a corpus can be determined quantitatively and qualitatively. Therefore statistical approaches of judging how good different corpus collection strategies are at providing good coverage are used. The methodology we adopt has been characterised by Leon (2005: 36/37) borrowing from Leech (1991: 106/107) thus:

- Focus on linguistic performance, rather than competence;
- Focus on linguistic description rather than linguistic universals;
- Focus on quantitative, as well as qualitative models of language;
- Focus on a more empiricist, rather than a rationalist view of scientific inquiry.

To carry out experiments we need the following:

i. First, one needs a language to work with. For this thesis we have selected the Setswana language.

ii. Second, one needs a corpus of such a language comprising samples of different text types on which experiments can be performed. For

experimentation, a 13 million-word Setswana corpus with a variety of text types on which experiments will be carried out has been compiled. The intended purpose of the corpus is defined as the aiding of Setswana dictionary compilation and research. While the corpus may be used for other kinds of linguistic research such as language variation and general linguistics, the corpus is primarily constructed for lexicographic purposes. Narrowing the purpose of the corpus to dictionary compilation and research is significant since it has implications on the kind of mark-up that needs to be undertaken on the corpus and the variety of text types that have to be included in the corpus design. The sampling of the intended corpus has been inspired by that of the BNC (Burnard, 1995 and Crowdy, 1991). The aim is to compile a synchronic corpus with texts from 1966 (post-Botswana independence). However, since texts covering broad varieties in Setswana are few, all texts have been considered for inclusion. The scarcity of texts in many categories, e.g. non-existence of newspapers[1], magazines, journals, and other printed matter in many African languages appears to have been a source of discouragement for tackling corpus design in many languages. The corpus that we have compiled is general, in that it is not restricted to any particular subject field, register or genre. Since its use is to test language for general language dictionaries, the corpus comprises a variety of text types from both spoken and written language.

iii. Third, one needs ways of determining how good a corpus is for lexicography. For this thesis we use keyword analysis, frequency lists and the measure of word types at 10,000 tokens intervals.

The statistical analysis is conducted by the use of a corpus querying software; WordSmith Tools (Scott, 2004-2006) which is an integrated suite of three main programs: wordlist, *Concord* and *Keywords*. The wordlist tool can be used to produce wordlists or word-cluster lists from a text and render the results alphabetically or by

[1] The *Naledi ya Botswana* newspaper which dates to the 1940s and *Mokgosi* newspaper of 2002-2005 have both seized distribution.

frequency order. It can also calculate word spread across a variety of texts. The concordancer, *Concord*, can give any word or phrase in context – so that one can study its co-text, i.e. see what other words occur in its vicinity. *KeyWords* calculates words which are key in a text i.e. used much more frequently or much less frequently in a given corpus than expected in terms of a general corpus of the language.

In our experiments keywords are first calculated for the different text types. Because of space constraints, the top 100 keywords of the test from each text type are given. The top 100 keywords constitute a limited version of the total results, however they are sufficient to advance and illustrate the line of argument we are pursuing. Second, type token measures of text types are calculated at comparable 10,000 token intervals. The aim is to determine lexical richness of text types at comparable points. The results shed a light on whether text types with a similar number of tokens have different word types. The significance of this experiment is in demonstrating that individual text types alone are limited in generating broad coverage word types which can be used generating a headword list. On the other hand, text types collectively complement each other in the word types they contribute. While certain text types may display a low number of types at 100,000 token intervals, such low types may be specialized and unique to the text type and therefore be valuable to the entire corpus.

Our argument is that for a corpus to represent a language, it must be designed in such a way that it includes a variety of text types from the language which it represents. The inclusion of such varieties of text types should be seen to be balanced. We discuss the subject of corpus balance and representativeness in Chapter 4. We will measure through keyword analysis if and to what extent different text types generate different keywords that are particular to them. The retrieval of unique word types from a text type gives support to the argument that a corpus that captures linguistic variability of a language community must be compiled using a variety of texts drawn from the text types of a language. Representing text variability in a corpus is significant since the quality of corpus-retrieved information for lexicographic purposes depends on the text input at the stage of corpus construction. This position finds support in Dash and Chaudhuri who argue that,

> The decision about what should belong to a corpus and how the selection is to

be made virtually controls every aspect of subsequent analysis. If designed methodically, it can reflect the language with all its features and qualities (Dash and Chaudhuri, 2000: 180).

1.5 Aims of the study

The aim of this thesis is to determine how Setswana corpora should be compiled and structured as balanced and representative entities through both quantitative and qualitative means in order for them to be "better suited" for lexicography. The aim is to measure whether a corpus compiled with texts from various text types or a corpus compiled with texts from few or a single text type generates words that are equally good for lexicography. We proceed from the assumption that text variability in corpus compilation is desirable. The assumption, however, demands empirical verification. Such verification can be achieved through experimentation which compares corpora and corpora components. To perform such comparisons accurately, we employ statistical methods since we agree with Kilgarriff (2000: 109) that "lexicographers need the skills and or the software to navigate through sometimes huge numbers of corpus instances." They need to apply statistical methods and natural language processing skills to make sense of the data. Such skills have been demonstrated in Bharathi et al. (2002). Bharathi et al., discuss the statistical analysis of ten Indian languages. The analysis is conducted using basic statistics like unigram frequencies, bigrams frequencies, syllable frequencies, word length distribution and sentence length distribution in the corpora of the ten languages. They were able to extract the following from the corpus (i) word frequencies and their percentages in the whole corpus (ii) the number of distinct words required to cover a certain percentage of corpus (iii) syllable frequencies and pattern extraction from syllables (iv) entropy of words in the corpus (v) word length analysis using average word length, modal word length and (vi) sentence length analysis using average sentence length, modal sentence length, etc.

The aim of this thesis is to determine if different text types contribute distinct word types. If this is found to be the case then such evidence would prove significant to corpus design for lexicography in general. The recognition that different text types

contribute different words, would then influence lexicographers, compiling dictionaries on the basis of corpus evidence, to pay particular attention to corpus design to ensure the broadest coverage possible of text types.

1.6 Research goals

In this thesis it is aimed to develop a model of corpus construction for the Setswana language which will provide a blue print for corpus design for languages similar to Setswana.

It is also the aim of this thesis to develop a structured Setswana corpus comprising a variety of text types to be used for experiments in this study and for future research of the Setswana language in size and context.

We aim to calculate and extract through keyword analysis words which are typical of different Setswana text types.

We aim to use frequency analysis to analyse and compare Setswana text types. Frequency will also be used to compare the Setswana corpus and the British National Corpus

We aim to measure and determine whether the representation of linguistic varieties in a corpus is crucial to a corpus output that reflects linguistic variability or whether similar outcomes may be achieved through building an archive of texts from a single genre.

1.7 Exposition of chapters

Following the introductory Chapter 1, the Setswana language is discussed in **Chapter 2**. In Chapter 2 the different contexts in which the Setswana language is used are examined. The different varieties of Setswana are relevant to corpus design, since what is modelled in a representative corpus is a corpus that reflects linguistic variability. We conclude the chapter by taking a historical view of Setswana research

in general and of the development of Setswana lexicography.

In **Chapter 3** we explore corpus lexicography by discussing what a corpus is and whether Web text qualifies as corpus material. Corpus applications on macro- and micro-structural levels are also discussed. We also introduce the exploration of corpora through frequency and keyword analysis and concordance lines inspection. The relevance of corpora to lexicography is discussed and we also examine some pre-electronic corpus studies and some early electronic corpus research. We conclude the chapter by reviewing a variety of methods of headword list identification and the previous use of corpora in Setswana dictionary compilation.

Chapter 4 explores a variety of issues in corpus design for lexicography. These are corpus balance and representativeness, corpus annotation, sample size, and spoken language in a corpus. These are followed by a discussion of how lexicographers have addressed the challenges of borrowing and code-switching in the Toqabaqita language and how their approach sheds light to the treatment of borrowings and code-switching in the Setswana language dictionaries. We conclude the chapter by reviewing the Brown Corpus and British National Corpus, illustrating their different strengths and weaknesses.

Chapter 5 discusses the Setswana corpus compiled during this study by examining texts included in the corpus components. The subcorpora types, tokens, type/token ratio (TTR) and standardized type/token ratio (STTR) are calculated.

Chapter 6 and **Chapter 7** are experiment chapters. In Chapter 6 we measure the different subcorpora through keyword analysis determining which words are typical of the various subcorpora. We demonstrate that different subcorpora are characterised by different keywords. In Chapter 7 we measure how for each text type the numbers of word types grow with every additional 10,000 tokens. The experiment is significant in that it measures types in a variety of text types at similar numerical intervals making it possible to make useful comparisons between the text types.

Chapter 8 concludes and summarises the findings of this study.

Chapter 2

The Setswana Language

2.1 The Botswana language situation

In this chapter the position of Setswana within a multilingual Botswana is discussed, situating it within a diverse national linguistic culture.

Botswana, a former British protectorate, is a landlocked southern African country. It has a population of about 1.7 million (2001 census)[2] in a land mass over twice the size of the United Kingdom (Botswana is 600, 370sq km while the United Kingdom is 244,820sq km)[3].

Botswana has an estimated 20 different languages spoken within her borders (Anderson & Janson, 1997: 7). Nyati-Ramahobo (1999: 80) estimates at least "22 distinct languages spoken in the country." These include amongst others: Khoisan languages (!Xoo, Nama, Kxoe!, Shua and others) Setswapong, Thimbukushu, Sekgalagadi, Shiyeyi, Otjiherero, Ikalanga, Setswana, English and many others. Despite its multicultural composition, only two languages, Setswana and English, occupy a dominant position in the educational setting (Mooko, 2004: 181/2). English is the official language and a language of considerable prestige, while Setswana, the language of the dominant Tswana peoples, is the national language and a lingua franca. Other Botswana languages apart from Setswana and English have no official status in Botswana (Molosiwa, 2004: 6) and remain excluded from functioning as mediums of instruction, excluded from being used in the media (both broadcast and

[2]The Republic of Botswana: Central Statistics Office, http://www.cso.gov.bw/

[3]The Central Intelligency Agency: The World Factbook: www.cia.com

print, save for Ikalanga which is used minimally in the *Mmegi* newspaper insert, *Naledi*), parliament, and in most public domains to communicate government policy. Minority languages are in general marginalised from any official function. However, in regions where they are the regionally dominant languages, for instance Mbukushu in north-western Botswana, they are usually used in official roles, like communicating with the chief or nurse (Hasselbring et. al., 2001: 32-33). Of the minority languages spoken in Botswana, Ikalanga is the language of the largest minority people. It is spoken mainly in the North-East and Central Districts of Botswana.

Table 1 gives the different language groups in Botswana and their associated ethnic groups together with regions where the majority of speakers are found. There is uncertainty over the exact number of people associated with different languages and dialects in the country. There are very few reliable figures on the sizes of ethnic groups and scholars at best give estimates of sizes of language communities (see Andersson and Janson, 2004, Hasselbring 2000, Hasselbring et. al., 2001). We therefore do not give any specific figures associated with the languages.

Table 1: Botswana's linguistic and ethnic structure

Linguistic Category	**Language Family Group**	**Associated Ethnic Groups**	**Administrative District**
SeTswana	Bantu, Southern	Bakgatla	Kgatleng
		Bakwena	Kweneng
		Bangwaketse	Southern: Ngwaketse
		Bangwato	Central
		Barolong	Southern: Barolong
		Batlokwa	South East
		Batawana	North West
		Balete	South East
		Bakhurutshe	Central
IKalanga	Bantu, Eastern	Bakalanga	Kgalagadi
Se-Birwa	Bantu, Southern	Babirwa	Kweneng,
Se-Tswapong	Bantu, Southern	Batswapong	North West
Se-Kgalagadi	Bantu, Southern	Bakgalagadi Bangologa Baboalongwe Bangologa Bashaga Baphaleng	Kgalagadi, Kweneng, North West
Shiyeyi	Bantu, Western?	Bayeyi	North West
Otjiherero	Bantu, Western	Baherero/Banderu	North West
Thimbukushu	Bantu,Western	Hambukushu	North West
Sesubiya	Bantu, Central	Basubiya/ Bekuhane	North West

Nama	Khoesan	Nama	Kgalagadi/Ghanzi
!Xoo	Khoesan, Southern	!Xoo	Kgalagadi & others
Ju/'hoan	Khoesan, Northern	Ju/'hoan	North West
Makaukau	Khoesan, Northern	Makaukau	Ghanzi
Naro	Khoesan, Central	Naro	Ghanzi
/Gwi	Khoesan, Central	/Gwi	Southern/Ghanzi
//Gana	Khoesan Central	//Gana	Central/Ghanzi
Kxoe	Khoesan, Central	Kxoe	North West
Shua	Khoesan, Central	Shua	Central
Tshwa	Khoesan, Central	Tshwa	Central/Kweneng
Afrikaans	Indo-European	Afrikaans	Ghanzi

Source: Selolwane (2004: 5).

Botswana's educational language policy of 1977 is a controversial document which does not recognize and encourage national linguistic diversity. It appears to be based on the belief that linguistic pluralism is a root source of ethnic and national unrest and not that it empowers citizens to meaningfully participate politically, socially and economically. Alidou (2004) has argued that in post-colonial Africa, in avoidance of ethnic wars, African governments ironically retained colonial languages which were viewed as neutral means of communication. She also argues that governments felt that in the interest of national unity, it was crucial that a country rallied behind a single flag, a single constitution and a single local language hence Setswana as a local language was adopted and sponsored by the Botswana government as a national unifying language. As Bagwasi (2003: 213) argues, "[t]he National Commission on Education 1977 states that Setswana is the language of national pride, unity and cultural pride." Alidou (2004) also observes rightly that in former British colonies African languages and English were used transitionally as medium of instruction and English became a dominant language after the fourth grade and the only language in secondary school and higher education. This state characterised by Alidou reflects the Botswana situation where the 1977 language policy entails the use of Setswana as the medium of instruction in standards (i.e. grades) 1 to 4, followed by a change-over to instruction in English from standard 5. A National Commission which reported in 1993 recommended a change in the policy so that English should become the medium of instruction right from the beginning of primary school, thus excluding Setswana from any such role. The government decided that (Republic of Botswana, 1994) instruction in Setswana is to be in the first year of primary education, and thereafter instruction had to be exclusively in English, save in the teaching of the Setswana language.

2.2 The Setswana language

Setswana is a member of a Sotho subgroup (also referred to as Sotho languages) of closely related Bantu languages found in southern Africa. This group includes Sesotho, spoken in Lesotho and certain parts of South Africa, and Sepedi, also known as Northern Sotho, which is spoken predominantly in the northern parts of Gauteng, around Pretoria in areas such as Polokwane in South Africa. Southern Sotho, Northern Sotho, and Setswana are largely inherently intelligible but have generally been considered separate languages (see also Cole, 1955: xv/xvi).

Setswana has mother-tongue speakers in at least four countries: South Africa, Botswana, Namibia and Zimbabwe. The largest number of speakers is found in South Africa (over 3 million speakers, about 8% of the population) where Setswana is one of the eleven official languages. Zimbabwe has an estimated 29,000 Setswana speakers and Namibia has approximately 6,000. In Botswana, Setswana is spoken by circa one million speakers (70-90% of the population) as a mother tongue (Andersson and Janson, 1997). Selolwane (2004: 4) observes that "…the SeTswana language is the most dominant of all the language groups found in Botswana, with at least 70% of the population identifying it as a mother tongue and another 20% using it as a second language." Seven percent speak other Sotho-Tswana languages (Setswapong and Sebirwa), 9% Ikalanga, 3% Seherero or Sembukushu, 2% Sesarwa (Khoisan), while 1% speaks Sesobea (Chikuhane) and 1% Seyei.

Her observations on the Setswana language are confirmed by Ramsay's (2006) report that 79% of Botswana's population speaks Setswana as a mother tongue. However other data varies considerably. Ramsay's data is from 2001 household census data.

Table 2: Number of speakers of Botswana languages

Language	Raw numbers	%
Setswana	1,253,080	78.2%
Ikalanga	126,952	07.9%
Sekgalagadi	44,706	03.5%
English	34,433	02.1%
Khoisan (Sarwa)	30,037	01.8%
Mbukhusu	27,653	01.7%

Sebirwa	11,633	00.7%
Chishona	11,308	00.7%
OtjiHerero	10,998	00.6%
SiNdebele	8,174	00.5%
Afrikaans	6,750	00.4%
Chikuhane (Subiya)	6,477	00.4%
Setswapong	5,382	00.3%
Seyei	4,801	00.3%
Nama (Sekgothu)	690	00.0%
Other African	10,036	00.6%
Indian langs.	1,848	00.1%
Other Asian	1,891	00.1%
Other European	804	00.0%
Other	864	00.0%
Unknown	3,368	00.2%

Source: Ramsay (2006) in *Mmegi* newspaper online (9th May 2006).

Ramsay's figures were however disputed by Nyathi-Ramahobo (Gaotlhobogwe, 2006) of Reteng[4] in *Mmegi* of Wednesday 10 May 2006. Reteng countered the data with its own estimates. It argued that unrecognized or minority tribes in the country number 1,030,000 or 60% of the total population, while the main tribes number 305,000 or 17.9% of the total population, with the rest (365,863 or 21%) consisting of immigrants. Reteng's data is speculative and cannot be trusted.

Literature on the language situation in Botswana usually makes a distinction between English as an official language and Setswana as a national language in Botswana. Setswana is seen generally as a language of national unity, and English as a language in which government policies are articulated (Arua and Magocha, 2002). This distinction in many instances is blurred with more of Setswana being used more in official contexts to explain government policies, which are written in English, and English encroaching into areas where traditionally Setswana has been used, such as funerals and weddings.

Setswana is a compulsory subject in Botswana government schools at both primary and secondary education (cf. Andersson and Janson, 1997: 21).

While in this thesis we devote greater focus to corpus development for the Setswana language in general, our focus will mainly be the Setswana language in Botswana,

[4] *Reteng* is a Botswana-based minority tribes' non-governmental organization.

and we will use Setswana language as used in South Africa for comparative purposes. Although Setswana has the largest number of speakers in South Africa, we choose to limit our research to Botswana where Setswana is spoken by the largest percentage of the population.

2.3 Setswana dialects

In Botswana, the majority of Setswana speakers are found in the Southern, Kweneng, and Central and North-West districts. Setswana has different regional dialects related to different tribal territories (see Table 1). The different Batswana tribal groups spread in Botswana "as a result of splits, secessions, and migrations" (Andersson and Janson, 1997: 22). The Bakwena are thought to have crossed into what is modern Botswana from northern South Africa around 1540. The Bangwaketse and Bangwato seceded from the Bakwena to form independent chiefdoms in the 17th century. In 1795 a group of Bangwato led by chief Tawana seceded and settled near Lake Ngami and gained control of north-western Botswana. The four Setswana dialects: Sengwaketse, Sekwena, Sengwato and Setawana are therefore related. The Sekgatla dialect spoken by the Bakgatla who live in and around Mochudi village in south-eastern Botswana is another dominant dialect which is associated with "standard" Setswana (Andersson and Janson, 1997: 27). There are other Setswana dialects spoken by other smaller Setswana tribes. These are Serolong, Selete, and Setlokwa. The larger part of the population of the country speaks the first four dialects (Sengwaketse, Sengwato, Sekwena, Sekgatla), which are numerically large. Setswana is generally used throughout the country as a lingua franca.

2.3.1 The village, cattlepost, lands and city language

On the construction of a spoken corpus, instead of looking just at the different social and regional dialects, there is also a need to be sensitive to the culture of the Batswana. Batswana have a complex way of living involving moving at different times between the lands (arable farms), the cattlepost (pastoral farms), the village and the city. This pattern of life cuts across tribal boundaries. It is significant to consider these four areas that characterise Batswana life since speakers across regional

varieties in these four areas tend to use language differently. In the city there is a great mixture of Setswana dialects and high levels of code switching between Setswana and English since there are greater levels of language contact and a greater concentration of educated people. The village has lower levels of language contact compared to the city, although it is more developed compared to the lands and cattlepost. It has distinct areas of Setswana usage like funeral and the *kgotla* (a traditional meeting place). The lands and cattlepost are usually inhabited by people who have never received any formal education, or if they have, it is minimal. They therefore use 'pure' Setswana and rarely code-switch and code-mix. They use basic utensils different from those in the city. There are no tarred roads, no electricity, no stoves, and the mode of transport is usually donkey carts or donkey backs, in most cases no tap water and many other things that characterise city life. The nature of discussions covers traditional issues; about rain and the lack of it; about the drought and complex names of plants and colours of animals. Their beliefs are different and they usually depend on traditional medicines and traditional beliefs. City and village dwellers that go to the cattlepost and lands usually adjust their speech to these environments. Recognising these differences would enhance the collection of diverse language usage and improve variability in texts collected for the analysis of Setswana.

2.4 Domains of Setswana language use

English dominates most of the written texts in Botswana and is used in tertiary education, even in the teaching of linguistics and literature classes at the University of Botswana, even though a Setswana workshop recommended "That the University of Botswana be approached and asked to teach Setswana in Setswana" (Moncho and Pandey, 1985: 33). Setswana remains the language of communication at home, social interactions in bars, sports, meetings in rural areas, funerals, public political meetings (*freedom squares*) churches and traditional meetings (*kgotla* meetings). Setswana is a national language and serves as a lingua franca (Bagwasi, 2003). Amongst the educated, there are great levels of code-mixing and code-switching, a subject we will revisit in Chapter 4.

2.4.1 Education

Instruction in government schools is in Setswana between standard 1 and 4 across all subjects, after which English is used as a medium of instruction. There is however a government move towards making all government schools 'English-medium' schools since it is believed that students with a good command of the English language perform better in their subjects. The Revised National Policy on Education (RNPE) (Republic of Botswana, 1994) recommends that "English should be used as the medium of instruction from Standard 2 as soon as practicable' (Rec. 18(a))" (Arua and Magocha, 2002: 450). Arthur's (1997: 230) research "demonstrates that an overwhelming majority of teachers reject the option of a Setswana-medium primary phase" while most teachers prefer English as "the sole medium of instruction throughout the primary school." Teachers therefore encourage students to use English inside and outside the classroom.

However, Setswana is frequently used for explaining difficult concepts through standard 7 and the first 2 years of secondary school. And it has been discovered that teacher-teacher and student-student interactions are always in Setswana (Nyati-Ramahobo, 1999: 131).

Setswana as a subject is compulsory from primary to the highest level of secondary education for all Batswana learners in government schools. A variety of texts are written in Setswana. We discuss these in section 2.5 of this chapter.

2.4.2 Setswana and media

Botswana has at least 10 newspapers[5], about 10 magazines and one government owned television station (Botswana Television (Btv)). There are four radio stations – two government owned and two private. Setswana is heavily used on the national radio station, *Radio Botswana***,** for interviews, news, live football broadcasts and general programming. Commercial radio stations like *Gabzfm*, *Yaronafm* and *RB2*

[5] *The Daily News, Mmegi, Monitor, The Botswana Gazette, The Botswana Guardian, The Tswana Times, Echo, The Voice, Midweek Sun, Sunday Standard*

broadcast almost exclusively in English.

On television Setswana is used for drama, news, debates, and sport broadcasts. Most magazines write exclusively in English and are imported from South Africa. Small parts of the government magazine, *Kutlwano***,** are in Setswana. These parts include stories and letters to the editor.

When we started this thesis there was one major Setswana newspaper, *Mokgosi*, established in 2002, which wrote exclusively in Setswana. The paper has since closed in 2005 because of lack of advertising and general disinterest of readers in news written in Setswana. *Mmegi*, the largest daily newspaper which writes mainly in English, has a two and a half pages Setswana insert called *Naledi*. The government owned daily, The *Daily News*, writes predominantly in English and has only one and a half pages in Setswana. Most Botswana newspapers write exclusively in English. These include amongst others *Monitor, Sunday Standard, The Midweek Sun, The Botswana Guardian, The Voice* and *the Botswana Gazette*.

2.4.3 The Courts

The Botswana legal system is made up of traditional and the common law courts (Nyati-Ramahobo, 1999: 86). The traditional courts, also known as customary courts, are presided over by a chief or his representative in a *kgotla* (a traditional meeting place). Proceedings are mainly carried out exclusively in the Setswana language. English is the official language of the magistrate court and the High court. While this is true, individuals can take an oath, plead, give evidence, verify facts or respond to court procedures in Setswana (Nyati-Ramahobo 1999: 88/9). Interpretation is usually offered in instances where those who appear before the court have minimum competency in English (Thekiso, 2001).

2.4.4 Parliament

English as the official language of Botswana is the main language for parliamentary debates. Although this is the case, members of parliament code-switch and code-mix

because of their multilingualism especially in English and Setswana.

2.4.5 Churches

Botswana's population is estimated to be 72% Christian[6]. The churches are diverse and follow different linguistic patterns. Hull (1987: 383) writing on the educational development in Botswana notes that, "formal education in most southern Africa was started by church missionaries.' It is therefore a matter of interest to study the linguistic situation of churches. The Zion Christian Church meetings are almost exclusively in Setswana while churches like the Seventh Day Adventist, The Anglican Church and the Roman Catholic Church use both Setswana and English for sermons, notices and songs. A similar pattern may be observed in various evangelical churches like Apostolic Faith Mission, Assemblies of God and Pentecostal Holiness Church where church notices and sermons are given either in English or Setswana with interpretations.

2.5 Text categories

In preceding paragraphs we have sketched contexts and areas of Setswana use. These areas are significant to corpus design in that they inform us of the text categories on which we can draw for the study of Setswana linguistic variability. Table 3 therefore gives a general outline of categories of texts in Setswana which could be compiled for the study of the language. The categories are listed in the general structure of the British National Corpus (Aston and Burnard, 1998).

Table 3: The Setswana text types rendered in the BNC style

Language	Usage types	Sources
Written Language		
Domain	Imaginative	Novels, short stories, poetry, plays, Popular lore
	Arts	Traditional Songs etc
	Belief and thought	Tracts, Bible, miscellaneous religious texts in other beliefs

[6] The Republic of Botswana: Central Statistics Office, http://www.cso.gov.bw/

	Commerce and finance	Business Manuals in Setswana
	Applied Science	Aids documents, TB literature, miscellaneous texts on clinical science
Medium	Book	Grammar texts, Botswana national: Vision 2016 text.
	Periodical	*Mokgosi* newspaper, *Naledi* newspaper and *Daily News*
	Misc. published	Survival International Text
	Misc. unpublished	Essays, letters etc
	To-be-spoken	Political Speech, Radio News Play text, Broadcast Scripts
Spoken Language		
Dialects & Region	Sekgatla	Kgatleng
	Sekwena	Kweneng
	Sengwaketse	Southern: Ngwaketse
	Sengwato	Central
	Serolong	Southern: Barolong
	Setlokwa	South East
	Setawana	North West
	Selete	South East
	Sekhurutshe	Central
Context Governed	Educational and Informative	Lectures talks, educational demonstrations, news commentaries, classroom interaction
	Business	Business meetings, trade union talks
	Public/Institutional	Political speeches, sermons, council meetings, Parliamentary Proceedings, court proceedings
	Leisure	Phone-ins, sports commentaries, club/society meetings
Interaction Type	Monologue	
	Dialogue	
	Unclassified	

2.6 Challenges of multilingualism and diglossia

Confronted with a language that does not have a long written tradition, corpus design and compilation presents unique challenges. Matters of balance and representativeness become difficult to maintain and define since the language is used in restricted areas. Scannell (2007: 2) has even argued that for such languages aiming for a representativeness corpus is absurd. Additionally, because of the bilingualism or

multilingualism of a speech community, code switching, borrowing and diglossia raise challenges that compilers of large corpora such as the BNC did not have to grapple with. Multilingualism matters are important in the construction of a Setswana corpus since Setswana historical contacts with Afrikaans and English have resulted with high levels of code switching and borrowing. For instance Cole (1955: 123) gives borrowing such as *keetane* and *galase* from *ketting* and *glas* (Afrikaans) and *buka* and *baesekele* from *book* and *bicycle* (English.)

2.7 The poverty of data

Section 2.5 discusses areas of Setswana use. The categories reveal the limited scope of the language use. While lexicographers working in Western languages have access to large amounts of electronic texts, for the construction of huge corpora running into millions of words of different genres covering newspapers, magazines, novels, academic texts, parliamentary pronouncements, and legal texts, African lexicographers work under great constraints because of the lack of data. Unlike their Western counterparts, they usually do not possess the luxury to be discriminative and selective of texts in electronic form since in the first place such texts are nonexistent. Many African countries do not use their indigenous languages in parliamentary debates, the publication of laws, instruction at schools and journalistic publications. This is certainly the situation in Botswana where there exists very little text in Setswana. In comparison with English, there are very few novels and plays in Setswana. There is also little instructional material in Setswana for lower primary school levels and virtually none for higher education. The only newspaper that wrote exclusively in Setswana, *Mokgosi*, closed down in 2005 because of lack of advertising and poor sales. One the papers which writes predominantly in English, *Mmegi,* also has a three and a half page Setswana insert, called *Naledi*. These low levels of written text give an idea of the gravity of the problem facing African lexicographers if they were to adopt the Western approach to corpus creation. They face practical constraints similar to those outlined by Rundell (1996) above, such as a shortage of time and money, the unavailability of machine-readable text, and copyright restrictions.

Although there are few written texts in African languages, their existence does not

guarantee that they are accessible to both native speakers and corpus researchers, or that the literate native speakers of the language read them. Many literate Africans rarely read texts in their own languages, although they may communicate extensively in such languages. The reason is not only because there is not enough written material in the African languages, but also because there is no culture of reading literature in African languages in many African communities. African lexicographers therefore face great hurdles in attempting to access both written and spoken texts for corpus construction. In cases where they have access to written texts, they run the risk of basing their research the attitudes of language purists and prescriptivists who remain wedded to a linguistic world that has never existed.

2.7.1 The Sanitised Data

Still on issues of written text, consideration needs to be given to the involvement of publishers and editors and the power of stylebooks on the written word, resulting in what can be called "sanitised data". Many publishers and editors have very rigid principles of which words should be used in their publications. They are heavily prescriptive, as in the newspaper *Mokgosi* which I worked for briefly. For example, the rare Setswana words *Mosupologo* (Monday), *Tshipi* (Sunday), *dira* (work, v.), and *kgwele* (ball) are generally preferred over the much more common *Mantaga*, *Sontaga, bereka,* and *bolo* respectively. Such preferences illustrate the biased prescriptive stance adopted by numerous publishers and editors who believe that borrowed language is not authentic and not part of the language. Their control of language does not reflect how the people use language, but rather reflects *how they wish it to be used.* A dependency on such language for the construction of corpora brings serious questions to the kind of corpora whose results have to be generalised to the entire language. This is especially so since corpora provide information about what to include and exclude, guides the lexicographer towards sharper sense distinction, and assists in selecting corpus-based examples (De Schryver and Prinsloo 2000b: 1). While "sanitised data" may be unavoidable, it is greatly unsatisfactory for dictionary research where generalisations about language use must be made. Instead, it should be considered together with spoken texts to obtain a clearer picture of the language use of a speech community.

2.8 Setswana language research

2.8.1 A historical overview

The known studies of the Setswana language may be traced as far back as November 1806 when the German, Hinrich Lichtenstein in *Ueber der Beetjuans* 'About the Batswana' (published in 1807), later translated into English (see Lichtenstein, 1973: 63), where he considered the various Batswana tribes as a single linguistic group and compiled what he referred to as the 'Beetjuana words'. He also lists in *Upon the Language of the Beetjuans* (1815: 478-488) a vocabulary of *The Beetjuan Language*. Around the same time, Henry Salt (1814: appendix, xxvii) records *A few words of the Mutshuana language copied from a manuscript journal of Mr Cowan*. The list includes the following words which we also render in current Setswana orthography with their English equivalents.

Table 4: Some of Henry Salt's Setswana terms

Salt's Setswana terms	Current orthography	English equivalent
let chãchi	*letsatsi*	sun
werri	*ngwedi*	moon
too na	*tona*	big, large, much
kom mo shu	*kamoso*	tomorrow

Campbell (1815: 221) also lists *Bootchuana Words* in his Travels.

Of great significance to the Setswana language is the Kuruman Mission station of 1824 with the expertise of Robert Moffat and his associates. In Kuruman in the LMS (London Missionary Society), Moffat rose to great significance, not only in the dissemination of Christian theology amongst the Batswana, but most importantly, and relevant to this chapter, in that he became the first person to reduce the Setswana language to a written form (Livingstone, 1857: 200).

The Setswana orthography was developed by the missionary Robert Moffat around 1820 and he based it on the Setlhaping dialect. The influence of the Setlhaping dialect has diminished and standard Setswana is now based on the Sekgatla, Selete, Sekwena, Sengwaketse and Sengwato dialects.

Moffat also translated the Bible and several hymns for his missionary expansion and in 1840 started training local converts to read the scriptures in Setswana so that they could propagate them amongst their own. Thus an interest in the Setswana language was mainly to "produce sound Christian teachers who [would be able to] preach the gospel, cope with white men, understand elementary business transactions and the value of land and evangelise Bechuana" (Moffat, 1842: 2). It was with the arrival of another missionary, Dr. David Livingstone, in 1841, in Kuruman that the education of locals increased and a school was built in Mabotsa in 1844. From then, there was an increase in Setswana research, most of it in the form of grammars books of the language. These amongst others, include works by, James Archbell's *A Grammar of the Bechuana Language* (1837), Rev. J. Fredoux *A Sketch of the Sechuana Grammar* (1864), A.J. Wookey (1904). These were later followed by more robust linguistic studies of the language, for instance the first Setswana phonemic study by Jones and Plaatje's (1916) and later Sandilands' (1953) *Introduction to Tswana* and Cole's (1955) *An Introduction to Tswana Grammar*.

2.8.2 The development of Setswana lexicography

In this section we trace the history of Setswana lexicography to the early missionary period and we situate it within missionary literacy programs amongst the Batswana. We then consider how developments in corpus and computational models have affected dictionary compilation and illustrate how the Setswana language could benefit from developments in corpora and corpus querying software (CQS) to produce frequency lists, concordances, and keyword analysis.

2.8.2.1 Lexicographic tradition

Setswana has a long lexicographic tradition characterised by low dictionary production. Jones (in Matumo 1993: vii) traces the origin of Setswana lexicography to John Brown's bilingual dictionary (1875), which is criticized by Kgasa and Tsonope (1998: iv) for its bilingualism, and to Robert Moffat's (1830) Setswana version of the Gospel of St Luke, which has definitions of difficult words in its final back pages.

> In 1830 Robert Moffat published a Setswana version of the gospel of St Luke, and at the back offered two pages of explanations of the more "difficult" words. Is it fanciful to regard this as the first small germ of a dictionary? …but the first published dictionary of which the Botswana Book Centre has record is that of John Brown in 1875 (Jones, in Matumo 1993: vii).

Cole (1955: xxviii) dates Setswana lexicographic research in later years in the plant names compilations of Miller (1951) and van Warmelo's (1931) lists of kinship terms.

However lexicographic research in Setswana dates much earlier than Moffat's 1830 writings that Jones refers to and certainly earlier than Cole's botanical and kinship references. Research demonstrates that Lichtenstein in the two volumes of *Travels in Southern Africa in the years 1803, 1804, 1805, and 1806* had a list of about 270 Setswana words and phrases. The original document in German appeared around 1811. Therefore the earliest lexicographic activity, at least of a headword list with its English equivalents, known to us so far can be traced to 1803-1806, in Lichtenstein works. In 1815, John Campbell in his *Travels in South Africa* gave a list of 80 'Bootchuana Words'. Salt (1814) in *Voyage to Abyssinia* contains a list of 20 *Mutshuana* words and their English equivalents. Therefore, lexicographical work in Setswana, regardless of its size and detail, existed before the work of Moffat, who came to Southern Africa in 1816.

The first published bilingual dictionary, *Lokwalo loa Mahuku a Secwana le Seeneles*, was compiled by John Brown (1875) of the London Missionary Society. An enlarged and revised version was published in 1895 and was reprinted in 1914 and 1921. In 1925 The Reverend John Tom Brown produced the third edition of this dictionary based on A.J. Wookey's research (Peters, 1982: xxiv). However since the 1925 dictionary version of Tom Brown to mid 1970s, no Setswana dictionary was compiled. It was not until 1976 that Morulaganyi Kgasa published his 134-page monolingual dictionary – *Thanodi ya Setswana ya Dikole* 'The Setswana Dictionary for Schools', whose main target group was primary school pupils. Kgasa's dictionary is the first Setswana monolingual dictionary published in Botswana. In 1998, in collaboration with Joseph Tsonope, Kgasa compiled the second monolingual

dictionary *Thanodi ya Setswana* which up to date remains the definitive monolingual Setswana dictionary. The dictionary used the Setswana standard orthography of 1981 (Ministry of Education, 1981). A smaller, but detailed, trilingual dictionary – Setswana, English and Afrikaans – was compiled by Snyman et al. (1990) whose target is the secondary school and university reader. *The Compact Setswana Dictionary* (1992) compiled by Dent is an abridged dictionary "intended for those people who find more comprehensive dictionaries too cumbersome or too detailed for their needs" (Dent, 1992: introduction). It has about 200, A6-sized pages. Matumo (1993) revised Brown's (1925) dictionary into what is now *Setswana-English-Setswana Dictionary*. Prinsloo (2004) reviews how this dictionary can be revised. The latest dictionary from Botswana is Créissels and Chebanne's (2000) *Dictionaaire Francais-Setswana Thanodi Sefora Setswana*, which is the first French/Setswana bilingual dictionary. Its primary target group is students of French at secondary and university level. It is the first and only Setswana dictionary with phonemic transcriptions and a large amount of pictorial illustrations. Cole (1995) has written *Setswana-Animals and Plants (Setswana-Ditshedi le ditlhare)* which is a dictionary of plants of animals although in the foreword of the dictionary, L.W. Lanham notes that "[t]he author of this remarkable book eschews the label "dictionary" for it, preferring to identify it as a "lesser listing of vocabulary" (Cole, 1995: ix). While Cole may disprefer the title "dictionary", his work is a bilingual dictionary, Setswana to English and English to Setswana and some of the entries are included with their Latin names.

2.9 Conclusion

This chapter laid a foundation for Chapter 5 which discusses Setswana corpus compilation and the two experiment chapters, Chapter 6 and 7. We have explored the varieties of Setswana and found out that Setswana's use is limited to certain domains. We also saw that Setswana lexicography may be traced as far back as November 1806 to the writings of Hinrich Lichtenstein in *Ueber der Beetjuans*. We also demonstrated that the first published bilingual dictionary, *Lokwalo loa Mahuku a Secwana le Seeneles*, was compiled by John Brown of the London Missionary Society in 1875. This chapter also identified Setswana dictionaries to give a picture of the degree of dictionary work in the language.

Chapter 3

Corpus Lexicography

> It is important to avoid perfectionism in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like (Sinclair, 2004).

3.1 Introduction

At the turn of the century Kilgarriff epically observed: "The arrival of electronic text corpora is causing a revolution in lexicography" (Kilgarriff, 2000: 109). Kilgarriff's statement rings true for many lexicographic projects in various languages which are aided by the exploitation of corpora. Of note is the contribution of the British National Corpus (BNC) to the production of Longman dictionaries (Summers, 1995) and many others and the effect of the Bank of English on the COBUILD dictionaries (Sinclair, 1996; De Beaugrande, 1997 and Moon, 2007). We start by defining what a corpus is, how it is of benefit to lexicography and we discuss two basic ways in which corpora are usually exploited.

3.2 What is a corpus?

What a corpus is, is usually characterised differently by various scholars.

Leech (1991: 8) defines a corpus as "a sufficiently large body of naturally occurring data of the language to be investigated". On the other hand, Renouf (1987: 1) defines a corpus as "a collection of texts, of written or spoken word, which is stored and processed on computer for the purpose of linguistic research". Sinclair (2004) defines a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." McEnery and Wilson (1996: 24)

define a corpus as "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration."

What recurs in these definitions is that a corpus is a language sample, a collection of texts, or pieces of language text for linguistic research. Kilgarriff and Grefenstette (2003: 334) however would see Renouf, Sinclair and McEnery and Wilson's definitions as characterized by "a smuggling of values into the criterion of corpus-hood" and conflating questions: "What is a corpus?" with "what is a good corpus?" They argue that "a corpus comprising the complete works of Jane Austen is not a sample, nor is it representative of anything else" and they define a corpus as "a collection of texts," a definition they qualify thus: "when considered as an object of language or literary study."

From the above definitions various points may be noted.

1. Corpora are usually "sufficiently large" for the research they have been compiled for. They usually run into thousands or millions of words (e.g. the 100 million-word British National Corpus). What "sufficiently large" translates to in terms of number of words or size of file is however not clear.
2. Corpora are collections of running texts. They are not just lists of words but rather chunks of texts like chapters of books, entire books, or transcribed speech.
3. Corpora are compiled for some linguistic research. "With a corpus stored in a computer, it is easy to find, sort and count items, either as a basis for linguistic description or for addressing language-related issues and problems" (Kennedy, 1998: 11).
4. Because of their massive size, corpora are usually stored in computers because of their storage and processing power. Cowie (1999: 117) observes that "nothing less than a computer revolution had taken place in lexicography." Kirkness (2004: 56) argues that "computers can store and process quantities of textual data quite unmanageable by humans" (see also Biber et. al. 1998: 22). Computers aid in querying corpora in fast and sophisticated ways (Kilgarriff and Grefenstette, 2003: 333/334) and modern corpora analysis and storage are characterised by a dependency on computers. Computers are "good at recall,

people are good at precision; that is, computers are good at finding a large set of possibilities, people are good judges of which possibilities are appropriate" (Kilgarriff, 2003: 1). Several advantages of the corpus-based approach emanate from the use of computers which make it possible to identify and analyse complex patterns of language use, allowing the storage and analysis of a larger database of natural language possible. Computers also provide consistent, reliable analysis. They can also "be used interactively, allowing the human analyst to make difficult linguistic judgements while the computer takes care of record-keeping" (Biber et. al., 1998: 4).

In this thesis we follow Cavagliá (2005: 5) and limit our definition of corpora to "language corpora" and exclude other media such as pictures and sounds.

Biber et al. (1998: 4) list four essential characteristics of corpus-based research as follows:

i. It is empirical, analyzing the actual patterns of use in natural texts;
ii. It utilizes a large and principled collection of natural texts known as a "corpus" as the basis of analysis;
iii. It makes extensive use of computers for analysis, using both automatic and interactive techniques;
iv. It depends on both quantitative and qualitative analytical techniques.

Biber (1995: 32) also lists the advantages of corpus based analysis as including:

1. The adequate representation of naturally occurring discourse, including representative text samples from each register. Thus, corpus-based analyses can be used on long passages from each text, and multiple texts from each register.
2. The adequate representation of the range of register variation in a language; that is, analyses can be based on a sampling of texts of a large number of spoken and written registers.
3. The (semi-)automatic linguistic processing of texts enabling analyses of much wider scope than otherwise feasible. With computational processing, it is

feasible to entertain a comprehensive linguistic characterisation of a text, analysing a wide range of linguistic features. Further, once the software tools are developed for this type of analysis, it is possible to process all available online texts.

4. Greater reliability and accuracy for quantitative analyses of linguistic features; that is computers do not get bored or tired – they will reliably count a linguistic feature in the same way every time it is encountered.
5. The possibility of cumulative results and accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be based on the same corpus of texts, or additional corpora can be analysed using the same computational techniques. Such studies can verify the results of previous research, and findings will be comparable across studies, building a cumulative linguistic description of the language.

3.3 Web as corpus

One of the alternative methods of corpus compilation is the construction of corpora from the Web (Jones and Ghani, 2000; Ghani et al., 2001). The Web currently contains billions of words. Kilgarriff and (2003) report of 172 million network addresses in January 2003. Fletcher (2002) points out that the Web has over ten billion publicly-accessible online documents which provide a comprehensive coverage of the major languages and language varieties, and span virtually all content domains and written text types. This massive language data is particularly available in major European languages like English, French, Spanish, Italian, Dutch and German. Smaller languages like Setswana however are underrepresented. In Chapter 5 of this thesis we discuss how together with Kevin Scannell of the Department of Mathematics and Computer Science, Saint Louis University, we compiled about half a million Setswana tokens using a Web crawler and downloaded web text for adding into the Setswana corpus used in this study.

At a theoretical level the question to ask is whether Web language text qualifies as corpus data. To this question, Kilgarriff and Greffenstette (2003: 343) answer in the affirmative. They respond to the charge that the Web is not representative by arguing

that "the web is not representative of anything else. But nor are other corpora, in any well-understood sense." We discuss corpus representativeness in considerable detail in Chapter 4.

De Schryver (2002) discusses the Web as and for corpus in African languages. He demonstrates that although the Web is highly dominated by English text, African languages are represented on the Web and can benefit from exploiting Web corpus in language research, such as spell checking and checking for grammatical patterns. Languages such as Swahili, Amharic, Hausa, Silozi and Chinyanja, isiZulu and isiXhosa have been demonstrated to exist in good numbers online.

The Web provides a cheap route to corpus compilation. An illustration of this is Ghani et al., (2001 and 2001a) who report on the CorpusBuilder architecture, query-generation methods and language filters of downloading documents for minority languages from the Web. By minority they refer to languages which are in the minority on the Web not necessarily a language spoken by a few people. CorpusBuilder works by taking as initial input from the user two sets of documents, relevant and non-relevant. Given these documents, it uses a term selection method to select words from the relevant and non-relevant documents to be used as inclusion and exclusion terms for query, respectively. The query is sent to a search engine and the highest ranking document is retrieved. This results with a large collection of text within a short time.

Fletcher (2005) gives the following points as support for using Web text:

- **Freshness and spontaneity:** the content of compiled corpora ages quickly, while texts on contemporary issues and authentic examples of current, non-standard, or emerging language usage thrive online.
- **Completeness and scope:** existing corpora may lack a text genre or content domain of interest, or else may not provide sufficient examples of an expression or construction easily located online; some very productive contemporary genres (blogs, wikis, discussion forums…) exist only on the Net.

- **Linguistic diversity:** languages and language varieties for which no corpora have been compiled are found online.
- **Cost and convenience:** the Web is virtually free, and desktop computers to retrieve and process web-pages are available to researchers and students alike.
- **Representativeness:** as the proportion of information, communication and entertainment delivered via the web grows, language on and off the Web increasingly reflects and enriches our language.

Baroni and Ueyana (2006) isolate the following advantages and disadvantages of using Web corpora. First advantages:

- Size**.** The Web has large amounts of text. Text size is important in NLP. Disambiguation algorithm performs better when trained on a larger amount of data (see also Bindi et al., 1994).
- The Web allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers.
- Web text can potentially contain a number of genres that are not present in traditional written sources such as blogs which generate vast amounts of spontaneously produced text.
- Web corpora tend to reflect more recent phases of a language than traditional corpora that are often subject to a certain lag between the time of production of the materials that end up in the corpus and the publication of the corpus.

They as well note the following disadvantages to Web corpora which are similar to those of any corpus built in a short time and with little resources.

- Web corpora are usually full of non-linguistic material and duplicated documents and duplicated text in different documents also referred to as ‘noise’.
- Since Web corpora are usually constructed with automated text mining methods, the researcher usually does not have full control over what ends up in the corpus, and cannot estimate the composition of the corpus.

- If a researcher plans to distribute a large Web corpus comprising millions of documents, (s)he will have a very hard time obtaining permission to use the documents from all the copyright holders.

Fletcher (2004) also finds the following disadvantages to Web-corpus data.

- It is difficult to establish authorship and provenance and to assess the reliability, representativeness and authoritativeness of texts since web-pages are typically anonymous and web server location is no certain guide to origin.
- Some sites have multilingual data.
- Other pages are authored by non-native speakers of varying competence, raising questions about language quality and influence of the source language.
- Certain longer prose text types predominate such as legal, journalistic, commercial and academic prose.
- Web text has no grammatical mark-up.

Some of the disadvantages listed by Fletcher are not unique to Web text. For instance, the argument that Web text has no grammatical mark-up is not limited to Web text since text from magazines and newspapers has no grammatical mark-up either.

Web text has been used for a variety of linguistic research. Amongst these is obtaining frequencies of bigrams unseen in a corpus (Keller et al., 2002). Shepherd and Watters (1998) propose what they term cybergenres and taxonomy of web-pages types and their evolution in an attempt to make sense of the structure of texts on the internet. Santini (2003) on the other hand proposes the development of computational methods to identify genres on the Web. Web text is therefore considered data that is suitable for linguistic analysis.

Electronic text corpora are valuable for language modelling in a variety of language technology applications such as speech recognition, optical character recognition, handwriting recognition, machine translation and spelling correction.

Kilgarriff (2001: 343) demonstrates that different researchers have used the web for a variety of research projects amongst these being:

1. As a source of language corpora for languages where electronic resources are in short supply.
2. As a source for bilingual parallel corpora.
3. To generate encyclopaedia entries.
4. For automatic distillation of lexical entries from empirical evidence.
5. In translation, translators when confronted with a rare term can find ample evidence of the term, its contexts, and associated vocabulary, through the simple use of a search engine.
6. The Web as a lexical resource, and as a source of test data, for Word Sense Disambiguation.
7. As a source for harvesting lists of named entities.

Similar benefits are discussed by Sharoff (2006).

The Web gives access to enormous quantities of text for free and it is still to be explored extensively in the study of Setswana. Chapter 5 reports on the use of a Web crawler to collect about half a million Setswana corpus from the Web. Below frequency profiling which will be used later in Chapter 6 and 7 is introduced.

3.4 Frequency profiling: frequency and type/token

3.4.1 Frequency counts

Frequency counts record the number of times each word occurs in a text. Sinclair (1991: 30) points out that "Anyone studying a text is likely to need to know how often each different word form occurs in it". This position is shared by Summers (1996: 261) that "all aspects of lexicography are influenced by frequency." Kilgarriff (1997: 135) furthermore notes that "A central fact about a word is how common it is. The more common it is, the more important it is to know it." Baroni (2006: 1) observes that "The frequency of words and other linguistic units play a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language." This

argument for frequency information is shared by Kilgarriff and Salkie who argue that:

> When a corpus is presented as a frequency list, much information is lost, but the tradeoff is an object that is susceptible to statistical processing. Word frequency lists are easy to generate, so measuring corpus similarity based on them will be viable in many circumstances where a more extensive analysis of the two corpora is not possible (Kilgarriff and Salkie, 1996: 121).

Baroni (2006: 3) observes that the data in a frequency list can be re-organized in two ways that are particularly useful to study word frequency distributions, namely as *rank/frequency profiles* and as *frequency spectra*. To obtain a rank/frequency profile, we simply replace the types in the frequency list with their frequency-based ranks, by assigning rank 1 to the most frequent type, rank 2 to the second most frequent word, etc. A frequency spectrum on the other hand is a list reporting how many word types in a frequency list have a certain frequency.

From our discussion in this section, it is clear that one of the basic corpus analyses is the frequency list, which reports a number of instances of each word type encountered in a corpus. Frequency counts therefore extract the different types of words, tokens or forms which make up a corpus.

A frequency list may be sorted in decreasing order from the highest ranking i.e. the most frequently used token down to hapax legomena (forms that occur only once in a given corpus) or vice versa. They are a "powerful tool in the lexicographer's arsenal of resources, allowing her to make informed linguistic decisions about how to frame the entry and analyse the lexical patterns associated with words in a more objective and consistent manner" (Summers, 1996: 266). We illustrate the use of frequency information in two English dictionaries *the Longman Dictionary of Contemporary English, 3rd edition* and *the Collins COBUILD Learner's Dictionary* in Section 3.5.

3.4.2 Type/token and word counts

In frequency analysis, there is a need to clarify what constitutes a *word* in a language and how words are counted. In linguistic literature the term *word* is defined in a

variety of ways. Some of these definitions while useful for theoretical linguistics, they are not very helpful in computational word counts. Finch (2000: 132) defines a word as "A unit of expression which native speakers intuitively recognise in both spoken and written language" and adds that "there is a certain indeterminacy about the definition of a word" (Finch, 2000: 132). Finch's definition is unhelpful in that "a unit of expression" could be anything from a word, a phrase, a clause or sentence. His definition also leaves the determination of what a word is to a speaker's intuition which may vary from one speaker to another. Aitchison (1992: 49) points out that "the best-known definition of a word is the one proposed by the American linguist Bloomfield who defined it as a minimum free form, that is, the smallest form that can occur by itself." She continues to argue that distinctions must be made between lexical items, syntactic words and phonological words. If we consider lexical items, the form such as *fly* represents at least two words:

fly N: an insect with two wings.
fly V: move through the air in a controlled manner.

The two lexical items have different syntactic forms associated with them. The insect could either be singular (*fly*) or plural (*flies*). The verb on the other hand could occur as *fly, flying, flies, flew, flown.*

Leech et al., (1982: 27) consider a word "delimited, for most purposes by a space (or punctuation mark other than a hyphen or apostrophe) on each side." However they also acknowledge that "the boundaries of words… are not always clear; e.g. we can write the sequence *piggy* + *bank* in three ways: *piggy bank*, *piggy-bank*, or *piggybank*."

In this study "a word is a minimal free form, the smallest unit that can exist on its own" (Dash and Chaudhuri, 2000: 189) and it is "delimited by a space…. on each side" (Leech et al., 1982: 27). From the brief discussions of what a word is it is clear that what a word is, is complex. Take for instance the following example:

> *His father will return from New York on Wednesday at 10am. Then they will stay in Pretoria for a week before buying a new house.*

One may observe that the above two sentences have 25 words: *His, father, will, return, from, New, York, on, Wednesday, at, 10am, Then, they, will, stay, in, Pretoria, for, a, week, before, buying, a, new* and *house.* Such a decision is arrived at by counting alpha-numeric characters delineated by spaces. Others may or may not consider whether digits and punctuation as words, a decision which will affect the shape of frequency distribution. Other decisions relate to token segmentation, whether it is *10 am* or *10am* or whether *New York* should be counted as a single token or two. Distinctions may also be made between upper and lower case forms. All these decisions will affect the number of elements counted. Distinctions between a word type and a token must be made so that there is no ambiguity and confusion concerning what is counted. Evert therefore emphasises that:

> For the purpose of obtaining frequency counts, it is essential to make a clear distinction between these two aspects: lexical items are called **types**, while their instances in a text are referred to as **tokens** (Evert, 2004: 33).

> A running word or a token is an arbitrary sequence (string) of letters delimited by spaces (Bergenholtz and Tarp, 1995: 34).

In a corpus of a 1,000 words the word form *kgomo* may occur 25 times. We say there are 25 tokens of *kgomo* which constitute a single word type. In this thesis, we follow Bergenholtz and Tarp (1995: 34) and assume words are separated by spaces and we count concords, grammatical words and numbers as distinct forms. A multi-word expression (MWE) will therefore be segmented where spaces exist and counted as multiple words. Words such as *gare ga bosigo* (midnight) or *bosigo gare* (midnight) will be counted as three and two distinct words respectively.

In African languages such as Setswana the matter of what a word is, is compounded by numerous concords in the language. While in English morphemes of verbs are suffixal and written conjunctively onto verbs, in Setswana the verbal prefixes (concords) are written disjunctively. This means that entities that are divided by white spaces are not always semantic words. Amongst these in Setswana are demonstrative concords, *lwa*, *jwa*; the quantitative concord, *ga*; relative concords, *sa*, *tsa*;

enumerative, *ga*, *wa*; nominal concords, *wa*, *ya*. The challenge could be illustrated with a text item such as *wa*. It could be a concord for first, second or third person singular of noun class 1 and noun class 1a. It could also be a concord for noun class three or eighteen. The use of the term 'word' in this thesis should therefore not be taken to mean that concords linguistically qualify as words and not morphemes. The use of the term word should rather be understood to mean graphical text items delineated by spaces. Such an approach will not be beyond criticism; however it does have some advantages to it; "the tradeoff is an object that is susceptible to statistical processing. (Kilgarriff and Salkie, 1996: 121).

3.5 Relevance of corpora to lexicography

Corpora are central to many lexicographic projects. De Schryver and Prinsloo (2000a: 292) note that "on the *macrostructural* level corpora provide crucial information for the creation of the lemma-sign list of dictionary, and on the *microstructural* level corpora enable lexicographers to tremendously enhance the accuracy of the dictionary articles themselves." On a macrostructural level they argue for lemmatised frequency lists and the need for lemma-sign distributions across sub-corpora to address typical macrostructural inconsistencies in many African-language dictionaries. They particularly expose the failure to include and treat commonly used words found in many other dictionaries as a lack of dependency on corpora. Their argument for the selection of lemma-lists on the basis of corpus frequency profiling is in line with Kilgarriff (1997: 136) who suggested the following processes were for improving the Longman Dictionary of Contemporary English, 3rd edition Summers, 1995) on the basis of frequency information:

- take a corpus
- extract a frequency list
- compare it with the dictionary to identify and rectify mistakes mismatches
- identify the one-, two-, and three- thousand cut-off points
- mark the corresponding dictionary entries accordingly

De Schryver and Prinsloo (2000: 298) argue not only that "frequency considerations

should determine the compilation of the lemma-sign list" but also that the lemma signs should be "in a sufficient variety of sources". This is what Leech et al. (2001: 17) refer to as word dispersion since it "is possible that the word has a high frequency not because it is widely used in the language as a whole but because it is "overused" in a much smaller number of texts, or parts of texts within the corpus." Scott (2004-2006: 123) refers to this phenomenon as consistency, that is, how consistent a word is used across a variety of texts of a corpus or subcorpus. Consistency analysis is calculated with every word frequency count in WordSmith Tools and is rendered in terms of the number of texts the word occurs in. This can be illustrated by considering the top 20 words in the Setswana corpus compiled for this study (see Chapter 5 for details). The top 20 tokens can be listed on the basis of decreasing frequency or on the basis of spread across texts (polytexty) as in Table 5 and Table 6. *Rank* refers to the position occupied by a token on a wordlist on the basis of its frequency. *Freq*. is the frequency of a word, or the number of times a word occurs in a corpus. *Texts* demonstrates the number of texts in which a word occurs. For instance "a" is ranked number 1 in Table 5 and it occurs 686,492 times in 3,055 texts.

Table 5: Top 20 words in the Setswana corpus ranked by word spread

Rank	Word	Freq.	Texts
1	a	686,492	3,055
2	go	418,088	3,000
3	e	413,176	3,016
4	le	358,736	2,977
5	o	336,417	2,990
6	ba	315,243	2,898
7	ka	290,557	2,956
8	ke	242,497	2,928
9	ya	228,511	2,959
10	mo	193,181	2,940
11	re	158,644	2,695
12	ga	149,529	2,851
13	fa	143,385	2,830
14	se	132,649	2,714
15	gore	125,686	2,828
16	di	124,651	2,807
17	ne	97,129	2,435
18	wa	94,822	2,803
19	tsa	92,885	2,772
20	sa	81,099	2,737

Table 6: Top 20 words in the Setswana corpus ranked by word spread

Rank	Word	Freq.	Texts
1	a	686,492	3,055
2	e	413,176	3,016
3	go	418,088	3,000
4	o	336,417	2,990
5	le	358,736	2,977
6	ya	228,511	2,959
7	ka	290,557	2,956
8	mo	193,181	2,940
9	ke	242,497	2,928
10	ba	315,243	2,898
11	ga	149,529	2,851
12	fa	143,385	2,830
13	gore	125,686	2,828
14	di	124,651	2,807
15	wa	94,822	2,803
16	tsa	92,885	2,772
17	sa	81,099	2,737
18	se	132,649	2,714
19	re	158,644	2,695
20	tse	69,238	2,640

Table 5 lists words on the basis of how frequent they are in the corpus. The most frequent word is ranked first, and the 11th frequent word, ranked 11, and so on. Table 6 on the other hand, lists words on the basis of spread, with a word found in the most number of texts ranked first. For instance, although *ba* is ranked 6th in terms of raw frequencies, it is ranked 10th in terms of spread. The rank of *ba* may be compared to that of *ya* which is ranked 9th in terms of raw frequency but is ranked in 6th in terms of spread.

Additionally, some of the words found in the raw frequency list do not make it into those listed in terms of spread. *ne* which occupies the 17th spot amongst the top 20 raw frequency list does not make it into the top 20 words on the basis of spread. Other words like *tse* which did not appear on the raw frequency list have been introduced into the list sorted by spread.

The two lists illustrate the fact that words that constitute a headword list should be abstracted on the basis of both raw frequency and word spread.

Corpora also have been used in the refinement of the dictionary microstructure, aiding in sense distinctions, the retrieval of typical collocations, frequent word clusters and the selection of authentic examples (De Schryver and Prinsloo, 2000a). De Schryver and Prinsloo (2000a) demonstrate that, particularly for African languages where this has been a weakness, corpora can tremendously aid to improve the quality of dictionary entries. Concordances, illustrative sentences in a dictionary and other linguistic information, which would be hard to generate through the use of non-corpus methods, are made readily available by the use of corpora and a corpus query system (CQS). For instance the sophisticated exploration of corpora by sketch engines (Kilgarriff and Rundell, 2002) has proved to be an efficient way of exploring the behaviour of words, the grammatical relations within which they participate, their collocational behaviour and thesaurus and "sketch differences" (which specify similarities and differences between near-synonyms) (Kilgarriff et al., 2004). A corpus therefore provides lexicographers with the information they need to compile authoritative descriptions of the vocabulary of a language. Lexicographers can retrieve the following from a corpus:

- **Statistical information**. From a corpus we can derive information about the relative frequency of different words or of the different grammatical constructions of the same word. This will reveal the inventory of the most common words which may be included as part of the dictionary's headword list, and the ones which are rare that may be left out of a dictionary. Thus Gomez (2002: 236) observes that frequency lists "enable lexicographers to take important decisions on which words a dictionary should include and which particular meanings". It is also possible to mark a word's frequency in dictionaries.

  We give examples of the marking of word frequencies in two English dictionaries: *the Longman Dictionary of Contemporary English, 3rd edition* (LDCE3) and *the Collins COBUILD Learner's Dictionary* (COBUILD).
  The LDCE3 was compiled using three corpora totalling over 135 million words: the 100 million words British national Corpus, the 30 million words Longman Lancaster Corpus and the 5 million word Longman Learner's Corpus. The dictionary marks the most frequent 6,000 words (3,000 entries from spoken transcribed text and another 3,000 from the written text) in the corpus on the page margins alongside entries. Spoken text is marked with S

and the written with W. The 3,000 words are further divided and marked by whether they are part of the top 1,000 spoken or written corpus (S1 or W1), the next thousand (S2 or W2) or the last thousand (S3 or W3) of the 3,000. For instance, the word **catch** is marked S1W1 to mean that it is part of the top 1 000 words in both spoken and written English. **Driver** on the other hand is marked S1W2 to mean that it is part of the 1,000 words of spoken English and it falls somewhere between the most frequent 2,000 and 3,000 words of written English. Such coding is essential since it guides a learner to words they are likely to meet and therefore need to learn.

Other frequency information in the dictionary, like meaning and homography, are not coded. The dictionary enters the most frequent meanings of a word first, and the less frequent ones later. For instance **chicken** has seven different meanings **1. ►BIRD◄ 2. ►MEAT◄ 3. ►SB WHO IS NOT BRAVE◄ 4. ►GAME◄ 5. which came first, chicken or egg? 6. a chicken and egg situation/problem/thing etc 7. your chickens have come to roost.** The meaning of *'bird'* is therefore more frequent in the language than *'somebody who is not brave'*.

Homographs are also shown in frequency order. The most common ones are entered and defined first while the less common ones are dealt with later. For example; **bound$_1$** (past tense of **bind**), **bound$_2$** (to be very likely to do…), **bound$_3$** (to run with a lot of energy) **bound$_4$** (noun, as in 'by leaps and bounds'). "by leaps and bounds" is rarer compared to **bound,** the past tense of **bind** in English. A learner would therefore be better off learning the past tense of **bind** before learning **bound** meaning "to run with lots of energy". Not only that, learners would be more likely to meet, in most texts, the most common meanings and if they look them up in a dictionary they would find them handled first, and not tucked in at the end. Such arrangement of senses is convenient since it ensures that words and meanings that students are likely to meet are arranged on the basis of their frequency.

The COBUILD gives frequency markers of entries to indicate how frequently they occur in the language. Instead of the S1 and W1 found in the LDCE3, they use a series of five diamonds ♦♦♦♦♦ in the extra column of the dictionary page. If all the diamonds are filled, then a word is one of the most frequent in the English language. The least frequent word has only one diamond filled ♦◊◊◊◊. There are nearly 700 entries representing 1,500

different forms which have five filled diamonds. The frequency of an entry includes that of its different forms (it is lemmatised), so that the frequency of the word *do* includes *does, doing, did,* and *done*. The next band of four filled diamonds ♦♦♦♦◊ covers over 1,000 entries which account for about 2,500 forms. Together with the five-filled diamonds band, the four filled diamonds words represent 75% of all common English usage. The 1,700 entries then represent essentially the core of the English language which is essential for a student to master. The next two bands of three black diamonds and two black diamonds ♦♦♦◊◊ cover a further 4,400 entries. The two filled diamond words ♦♦◊◊◊ include such words as *shuttle, shy, sickness, shrub, shrink, mounted, minimal, minus, midst,* and *soap*. Entries with a single black diamond ♦◊◊◊◊ represent the rare but important words which might have a restricted context of usage, they may be literary or words with specialized usage. The back matter of the dictionary comprises over 3,000 entries (Sinclair, 1996:1316-1322) accounting for nearly 10,000 forms. The decision to list them as part of the back matter is useful since students can assess their vocabulary power by simply reading the list and identifying those words that are unfamiliar to them and then finding their meanings in the dictionary. Teachers too can use the wordlist as a basis of class exercises to teach learners the core English vocabulary.

- **Alternative forms and spellings**. The corpus, if it is large enough, should present alternative spellings/forms of words (e.g. program/programme) and facilitate judgements as to which form should be used for the primary spelling. It will also reveal which forms are common enough to need to be entered as cross-references.

- **Semantic information**. From a corpus lexicographers can extract evidence for the word's different meanings and nuances. If the corpus is large enough, it would be able to show the most common senses of the word, and suggest the order in which such senses should be listed in a dictionary. It should also demonstrate common applications of the word (Biber et al., 1998).

- **Collocations**. From a corpus lexicographers can extract concordance lines to study the company that words keep. The computer concordance will reveal the most common collocations for individual words, by the words which tend to

come immediately before and after it in the concordance printout. It will assist describing the collocations which support particular senses, and their relative frequency would help to decide the order in which they should be listed in a dictionary. Collocation analysis will also lead to the isolation of idiomatic expressions (see Section 3.10).

- **Typical stylistic contextual information**. A large corpus with written and spoken data can provide examples from a variety of texts and sources, and thus indicate whether a particular word is current across the stylistic scale, or confined to spoken or written text only. In dictionaries such information may be turned into stylistic labels, such as "colloquial", or usage notes "found mostly in spoken interaction". Baugh et al. (1996: 44) argues that a corpus is critical in providing *context of use* information in the following areas: register (formal, informal, slang, taboo, taboo slang etc.), special context of use (specialised, medical, law, literary, poetic), language variety (American, British, and Australian), general context of use, e.g., speaker attitude (approving, disapproving).

- **Examples of actual use of the word**. Authentic examples from the corpus help to show the grammar and semantic functions of words, to remind L1 readers of the patterns of usage, and teach them to L2 readers. The dictionary-maker then has to decide whether such examples should be found for all words (including more technical ones), and for all senses of a word. The contents of the corpus example may therefore be needed to complement the definitions. Sinclair (1991: 39) argues that the “initial evidence should always be… from the observation of language in use”.

Baugh et al. (1996: 44) additionally argues that a corpus is beneficial to dictionary compilation in that it demonstrates the typical subjects and/or objects of a verb (see Section 4.3 of Chapter 4) and reveals encyclopaedic information of a word.

The contribution of a corpus to the dictionary making process has been discussed extensively in lexicographic literature by Béjoint (2000: 97), Sinclair (1987) and Sinclair (1991) who demonstrate the crucial nature of a corpus to the dictionary

compilation process.

3.6 Some pre-electronic frequency studies

Studies of frequency lists culled from corpora are not a recent occurrence and predate corpus computational developments. Kennedy (1998: 13-19) reports on pre-electronic corpora before the 1960's in five main fields of scholarship which we summarise below:

1. **Biblical and literary studies:** From at least the 18th century the Bible as a corpus has been used to generate lists and concordances to show that the Bible parts were factually consistent with each other. An example of such work is Cruden's concordance of 1736.

2. **Lexicography:** Corpus lexicographic work may be traced to 17th century Samuel Johnson's large corpus of sentences from writers to illustrate meanings and uses of English words (Sinclair 1991: 40). The compilation of the *Oxford English Dictionary* (OED) by James Murray and associates was also corpus-based; with over 2000 readers collecting millions of citations to illustrate word usage. In America, Noah Webster compiled *An American Dictionary of the English Language* in 1828 with the help of citation slips comprising millions of words (Kennedy, 1998: 14).

3. **Dialect studies:** Nineteenth century linguists compiled corpora to explore lexical variation in the choice of words for particular concepts (Kennedy, 1998: 14/15).

4. **Language education studies:** Thorndike (1921) compiled a 4.5 million-word corpus from 41 sources and generated a frequency list to aid curricula materials for teaching. J.W. Kaeding with the aid of assistants developed an 11 million word German corpus to gather statistical information of German words and letters to improve the training of stenographers.

5. **Grammatical:** Other corpora have been compiled to be used as sources of descriptive grammars. Among these is the work of Jespersen (1909-49).

3.7 Electronic-corpora studies

The first electronic corpus was the one million *Brown University Standard Corpus of Present-Day American English* commonly known as the Brown Corpus by Francis and Kučera (1964). It comprised 500 samples of 2000 words of continuous written English.

A similar corpus to the Brown Corpus, the Lancaster-Oslo/Bergen (LOB) Corpus, was compiled in the late 70s to study British English (Johansson and Hofland, 1989) through frequency analysis. Recently mark-up and word frequency studies have been done on the 100 million-word BNC (Leech et. al., 2001 and Rayson et al., 2002). These detailed studies do not only list alphabetical and rank frequency lists of the whole corpus, but include frequency lists of spoken versus written parts of the corpus, and unlemmatized frequency lists of spoken and written parts of the BNC. The studies investigate the frequency lists of the demographically sampled and context governed part of the spoken BNC.

3.7.1 An example of frequency profiling

Below the value of frequency profiling is illustrated by studying words which are characteristic of a particular genre. We look at the sports and business text. For our experiment parts of about a million tokens of the *Mokgosi* newspaper text are used. *Mokgosi* was a Botswana newspaper that wrote exclusively in the Setswana language. It closed down in 2005. The *Mokgosi* newspaper text is divided into five categories, the number of tokens given in brackets: Arts & Culture (476,523), News (1,426,223), Letters (502,729), Sport (289,205) and Business (247,246). For this part we are only interested in Sport (289,205) and Business (247,246) subcorpora from which we generate frequency counts using WordSmith Tools version 4.0 (Scott, 2004-2006). From our results we give the top 100 words for each including functional words and then offer the results again of the top 100 words, with functional words excluded.

In Table 7 below *N* is the number a word occupies in the list in terms of its frequency, this is the same as its rank. *Freq*. is the frequency of a word, or the number of times a word occurs in a corpus.

Table 7: Top 100 Mokgosi sport tokens with functional words

N	Word	Freq.
1	A	11,596
2	E	10,999
3	o	10,175
4	go	6,682
5	ba	6,566
6	le	6,129
7	ka	4,397
8	ya	3,835
9	mo	2,569
10	fa	2,315
11	ke	2,253
12	se	2,162
13	re	2,117
14	gore	2,085
15	di	1,890
16	ga	1,850
17	ne	1,742
18	sa	1,724
19	wa	1,600
20	kwa	1,559
21	tsa	1,238
22	tse	1,177
23	la	1,120
24	tla	1,118
25	bo	893
26	mme	748
27	bone	739
28	fela	721
29	setlhopha	649
30	jaaka	552
31	nna	532
32	batshameki	520
33	jwa	489
34	motshameko	473
35	na	464
36	yo	461
37	ngwaga	454
38	aforika	445
39	jalo	429
40	neng	406
41	morago	402
42	metshameko	367
43	ditlhopha	357
44	kgaisanyo	356
45	gagwe	348
46	dira	328
47	lekgotla	324
48	madi	318
49	botswana	310
50	thata	283
51	lefatshe	279
52	borwa	275
53	ene	267
54	ntse	267
55	pele	266
56	tswa	264
57	mokgosi	258
58	bona	256
59	kgwele	247
60	teng	247
61	batho	243
62	nako	243
63	mafatshe	242
64	bangwe	235
65	bobedi	235
66	ntlha	232
67	mongwe	231
68	masome	225
69	sentle	222
70	tlhalositse	221
71	eo	210
72	yone	200
73	setse	197
74	kgwedi	189
75	simolola	189
76	dipitse	188
77	motho	188
78	gone	186
79	kgaisano	180
80	mono	180
81	gape	179
82	tshameka	178
83	dinao	176
84	rona	176
85	bile	169
86	jaanong	169
87	setshaba	168
88	komiti	165
89	tsenelela	160
90	itse	159
91	mabedi	158
92	lesome	154
93	tota	154
94	tshwanetse	152
95	sena	151
96	bomme	149
97	santse	148
98	tiro	148
99	tlhalosa	147
100	batla	145

Table 7 reveals that *a, e, o, go, ba, le, ka, ya, mo, fa*, are the top ten most frequent words in the sport subcorpus. These words are members of the closed word classes (also known as function or grammatical words) which include classes such as concords, pronouns, and numerals (Leech et al., 1982). At least 35% of the words in

Table 7 are functional words. We find that the first 28 words are all functional words. It is common to most frequency lists to have functional words at the top of frequency lists.

Therefore in determining the frequency of the most frequent tokens in a corpus it may be attractive to remove the functional words from the list. Since functional words are usually the most frequent in corpora, they may in certain cases not provide critically comparative information between lists. Removing them from wordlists and remaining with content words (open-class words) may aid lexical comparison between lists in certain cases. This argument is not new. Gomez has argued before for English analysis that:

> The main problem with this information (frequency list of functional words) is that the use of raw frequencies highlights the very common words such as *the, of, in etc.,* despite the fact that their comparatively high frequencies of occurrence are unlikely to provide conclusive evidence of any specifically used vocabulary in any sublanguage (or corpus). These are words that, on the basis of frequency of occurrence alone, would be found to occur within most sublanguages, and it can perhaps be read more usefully if the purely grammatical words (close-word items) are discarded (Gomez, 2002: 239).

The argument is therefore that the top 100 words would be read informatively if the functional words were discarded from the list. Their removal would reveal content words that could define a genre and provide comparative information. We therefore removed functional words from the *Mokgosi* Sport wordlist's top 100 tokens. The top 100 words excluding functional words are:

Table 8: Mokgosi sport list's top 100 tokens without functional words

N	Word	Freq.	8	neng	406	16	madi	318
1	Setlhopha	649	9	morago	402	17	botswana	310
2	nna	532	10	metshameko	367	18	thata	283
3	batshameki	520	11	ditlhopha	357	19	lefatshe	279
4	motshameko	473	12	kgaisanyo	356	20	borwa	275
5	ngwaga	454	13	gagwe	348	21	ntse	267
6	aforika	445	14	dira	328	22	pele	266
7	jalo	429	15	lekgotla	324	23	tswa	264

24	mokgosi	258
25	bona	256
26	kgwele	247
27	teng	247
28	batho	243
29	nako	243
30	mafatshe	242
31	bobedi	235
32	ntlha	232
33	mongwe	231
34	masome	225
35	sentle	222
36	tlhalositse	221
37	setse	197
38	kgwedi	189
39	simolola	189
40	dipitse	188
41	motho	188
42	kgaisano	180
43	mono	180
44	gape	179
45	tshameka	178
46	dinao	176
47	bile	169
48	jaanong	169
49	setšhaba	168

50	komiti	165
51	tsenelela	160
52	itse	159
53	mabedi	158
54	lesome	154
55	Tota	154
56	tshwanetse	152
57	sena	151
58	bomme	149
59	santse	148
60	tiro	148
61	tlhalosa	147
62	batla	145
63	seka	145
64	tshwana	145
65	nngwe	144
66	fetileng	143
67	dilo	142
68	fenya	141
69	jang	140
70	batswana	139
71	gompieno	139
72	rre	138
73	kgang	137
74	motshameki	137
75	sengwe	135

76	dingwe	133
77	tlang	133
78	gae	132
79	nno	130
80	basimane	129
81	maemo	129
82	ise	127
83	bontsi	126
84	liki	125
85	metshamekong	120
86	tsile	120
87	jaana	119
88	nne	118
89	pedi	116
90	tshameko	114
91	morule	113
92	bfa	112
93	gaborone	112
94	mathata	112
95	tsaya	112
96	tsena	112
97	bnsc	110
98	boletse	110
99	tlase	110
100	dikgaisanyo	109

The list of content words reveals clearly the genre of sport through the use of the following words *setlhopha* (team) (1), *batshameki* (players) (3), *motshameko* (game) (4), *metshameko* (games) (10), *ditlhopha* (teams) (11), *kgaisanyo* (competition) (12), *madi* (money) (16), *nako* (time) (29), *simolola* (start) (39), *Dipitse* (Zebras – a nickname for the Botswana football team) (40), *kgaisano* (competition) (42), *setšhaba* (nation) (49), *basimane* (boys) (80), *liki* (league) (84), *tshameko* (play, noun) (90), *pedi* (two) (89), *BFA* (Botswana Football Association) (92).

The frequency list has helped isolate the most common words that are characteristic of the genre purely on the basis of their frequency. However other words in the top 100 wordlist are not distinctive to the genre. Such words include *tshwana* (same as), *tlhalosa* (explain), *tota* (truly), *fetileng* (past), *ise* (has not), *boletse* (told/said), *jalo* (like that), *mathata* (problems), *Morule* (December), *jaanong* (now), *dilo* (things), *maemo* (positions), *tsaya* (take), *batla* (want/seek), and a few others. This is not surprising since the top 100 words are raw frequency outputs and are not isolated on any measure that isolates words which are typical to, or stand out in, a text.

3.8 Keyword analysis

A much more precise method of identifying words particular to a genre is through the calculation of keyness which isolates words which are "key" to a corpus or subcorpus since these are useful in characterising a text or genre. We will implement our calculations by using a KeyWord tool which is part of WordSmith Tools version 4 (Scott, 2004-2006). The program has been used previously successfully for comparing corpora (Berber-Sardinha, 2000; Scott, 1997, Xian and McEnery, 2005).

To conduct the calculations, two corpora or subcorpora are required: one large another small. The large one is used as a reference file, while the small one is the study corpus, the one we are interested in studying. A reference corpus has been referred to as a "'normative corpus' since it provides a text norm (or general language standard) against which we can compare" (Rayson et al., 2004: 2). Two wordlists are generated from the two corpora. The aim is to find out which words characterise the text that is analysed. Keyness is "calculated by comparing the frequency of each word in the wordlist of the text you're interested in with the frequency of the same word in the reference wordlist" (Scott, 2004-2006: 92). The result is a list of keywords, or words whose frequencies are statistically higher in the study corpus than in the reference corpus. These are known as *positive keywords*. The software also identifies words whose frequencies are statistically lower in the study corpus. These are called *negative keywords*. In this study it is the positive keywords that we are interested in i.e. words occurring with a higher frequency than expected.

For this experiment the study corpus is the *Mokgosi* Sport section with 289,205 tokens and we compare it against our reference corpus, for which we will use the *Mokgosi* News section text which has 1,426,223 tokens. We provide the results below of only the top 100 most frequent tokens.

Table 9: Mokgosi top 100 sports keywords

N	**Keyword**	2	setlhopha	4	kgaisanyo
1	batshameki	3	motshameko	5	ditlhopha

6	metshameko
7	kgwele
8	kgaisano
9	dipitse
10	aforika
11	motshameki
12	tshameka
13	liki
14	bfa
15	bnsc
16	tshameko
17	mokatisi
18	kgaisanong
19	thenese
20	metshamekong
21	tla
22	dikgaisanyo
23	sofotebolo
24	nno
25	tsenelela
26	sejana
27	kgaisanyong
28	morule
29	zone
30	dinno
31	popa
32	basimane
33	setlhopheng
34	ngwaga
35	volleyball
36	motshamekong
37	karate

38	ketlogetswe
39	ikatisa
40	ikatiso
41	dinao
42	boramabole
43	dikgwele
44	oosi
45	bdf
46	batabogi
47	nosa
48	notwane
49	mabelo
50	mokatise
51	veselin
52	mokganedi
53	rollers
54	bobedi
55	fenya
56	lefela
57	tunisia
58	lebaleng
59	iponela
60	Diolimpiki
61	mabole
62	dipetsana
63	fighters
64	dietsele
65	fani
66	morocco
67	motshwara
68	liking
69	marumo

70	keatlholetswe
71	kemoeng
72	molefhe
73	soobolo
74	luza
75	tlhaodi
76	netebolo
77	borwa
78	nigeria
79	komiti
80	cosafa
81	bakatise
82	motsotsong
83	botsamaise
84	ditlhopheng
85	tafic
86	mono
87	championships
88	mafolofolo
89	kutlwano
90	libya
91	dikgaisano
92	ikatisong
93	molwantwa
94	karolong
95	dikgaisanyong
96	fifa
97	phenyo
98	kirikete
99	rugby
100	tshamekile

The above *Mokgosi* sports 100 keywords offer us a better streamlined list of terms that are key in the genre of sports. This list is more precise than a list generated on the basis of frequency. It isolates those terms which are only key to the genre from the corpus. The list includes names of **teams** like *Dipitse* (9), *Popa* (31), *BDF* (45), *Notwane* (48), *Rollers* (53), *Tunisia* (57), *Fighters* (63), *Nigeria* (78), *Tafic* (85), *Mafolofolo* (88), *Kutlwano* (89); names of different **sports/games**: *kgwele* (football) (7), *thenese* (tennis) (19), *sofotebolo*/*soobolo* (softball) (23/73), volleyball (35), karate (37), *netebolo* (netball) (76), *kirikete* (cricket) (98), rugby (99); names of **sports associations and organisations**: BFA (14), BNSC (15), COSAFA (80), FIFA (96); names of **sport personalities**: (*Tom*) *Ketlogetswe* (the name of a sports journalist) (38), *Veselin* (the name of the former Botswana national football team coach) (51), *Mokganedi* (the name of a sports journalist) (52), *Marumo* (the name of a footballer)

(69), *Kemoeng* (the name of Botswana National Sports Council Chairperson) (71), *Molefhe* (the name of an athlete) (72), *Tlhaodi* (The assistant chairperson of Botswana Tennis Association) (75), *Molwantwa* (the name of a footballer) (93), *Luza* (the name of a boxer) (74), and many others. The list also includes **sport verbs** amongst these being *tshamekile* (played) (100), *tsenelela* (take part in/attend) (25), *ikatisa* (train) (39), *fenya* (win) (55), *iponela* (got/won) (59). The list includes **sport positions** amongst these being *komiti* (committee) (79), *botsamaise* (leadership) (83), *motshameki/batshameki* (player/players) (11/1), *mokatise* (coach) (50). That the top 100 most key tokens contain data from diverse games including names of officials and players suggests that keyness analysis is crucial for isolating data that is particular to a genre.

3.9 Business keywords

The keyword analysis experiment was repeated to extract business keywords. For this experiment the study corpus is the *Mokgosi* Business section with 247,246 tokens and it is compared against our reference corpus which is again the *Mokgosi* News section text which has 1,426,223 tokens.

Table 10: Mokgosi business keywords

N	keyword
1	ditlhaeletsanyo
2	kgwebo
3	ndlovu
4	thapelo
5	ditshupo
6	banka
7	penrich
8	kompone
9	boripana
10	dikgwebo
11	itholo
12	dibanka
13	bagwebi
14	letlole
15	dikompone
16	bojanala
17	dibonto
18	bedia
19	bobs
20	agoa
21	aforika
22	peepa
23	dithoto
24	mafatshe
25	koporase
26	yuropa
27	khemikhali
28	lekalana
29	air
30	letlalo
31	madi
32	funeral
33	kgwebong
34	provida
35	thulaganyo
36	dipesente
37	ditlamelo
38	ditlhotlhwa
39	koafatsa
40	botlhole
41	boccim
42	lenaneo
43	itsholelo
44	mookamedi
45	reka
46	alafasegeng
47	beci
48	kelebogile
49	tlhalosa
50	foxcroft
51	hemilwe
52	hemiwa
53	sacu
54	lenchwe
55	diselula
56	tsogwane
57	sekoloto
58	lefhenya
59	rialo

60	difofane
61	lenaneong
62	kotsi
63	mhama
64	mmaraka
65	barclays
66	tlhong
67	makasine
68	mokgopha
69	peugeot
70	privatisation
71	siwawa
72	solofelwa
73	thekiso

74	dithentara
75	diphatsa
76	taolo
77	bota
78	galeforolwe
79	dikoketso
80	jab
81	tlhotlhwa
82	ceda
83	allan
84	batlamedi
85	diresiti
86	golden
87	kgokagano

88	nshakazhogwe
89	okacom
90	ppadb
91	privatization
92	thema
93	boranyane
94	mono
95	mabenkele
96	diaparo
97	boherabongwe
98	orange
99	papadisanyo
100	tlaabo

The business text in the *Mokgosi* newspaper is written by the business journalist, Thapelo Ndlovu. His surname and first name occupy position 3 and 4 respectively in the above list. The list also includes names of the following **companies, businesses and organizations**: *Penrich* (7), *BEDIA* (18), *BOBS* (19), *AGOA* (20), *BOCCIM* (41), *BECI* (47), *SACU* (53), *Barclays* (65), *Peugeot* (69), *BOTA* (77), *JAB* (80), *Allan* (for Allan Gray) (83), *OKACOM* (89), *Orange* (98); **business nouns** *ditlhaeletsanyo* (communications) (1), *kgwebo* (business) (2), *kompone* (company) (8), *letlole* (saving) (14), *dikompone* (companies) (15), *dibonto* (bonds) (17), *dithoto* (goods) (23), koporase (corporation) (25), *lekalana* (department/sector) (28), *madi* (money) (31), *itsholelo* (economy) (43), *sekoloto* (debt) (57), *mhama* (sector) (63), *dithekiso* (sales) (73), *dithentara* (tenders) (74), *diresiti* (receipts) (85), *boranyane* (technology) (93), *mabentlele* (shops) (95), *diaparo* (clothes) (96), *papadisanyo* (trade) (99) and many other terms; **business personalities**: *Thapelo Ndlovu* (business journalist) (2/3), *Kelebogile* (*Rantsetse*, the name of a local entrepreneur) (48), (*Slumber*) *Tsogwane* (Assistant Minister of Finance and Development Planning) (56), *(chief) Lenchwe* (54), (*Kagiso*) *Lefhenya* (young entrepreneur) (58), (*Tshidi*) *Tlhong* (chairperson of Junior Achievement Botswana) (66), *Mokgopha* (the name of a cobbler) (68), (Anthony) *Siwawa* (chairperson of CVF) (71), (*Joshua*) *Galeforolwe* (chairman of PEEPA) (78), (*Ishmael*) *Nshakazhogwe* (chairperson of Zambezi Motors) (88).

The relevance of these lists is to be evidence for the power of frequency profiling in lexical analysis. Frequency counts assist in the identification of most significant words on the basis of their frequency. Keyword analysis aids the retrieval of different genre-specific terms. When such analyses are repeated on texts from different genres and text

types, we would end up with keyword lists from different genres of a language, which could be combined to provide a broad vocabulary of such a language. For lexicography, the challenge with compiling a corpus with different text types of different sizes and then studying the raw frequency lists of the entire corpus together is that such an approach may obscure the keywords in various corpus components since certain keywords from a particular text type may be pushed lower in the frequency list and therefore risk exclusion. Studying different genres in isolation can aid in ensuring that different genres are represented and reflected in the dictionary headword list. The results may also be crucial in aiding marking entries as frequent in certain genres.

In Chapter 6 and 7 we use frequency profiling to measure lexical density and isolate keywords from a variety of genres in the Setswana corpus. Such experiments will be aimed at proving that subcorpora are characterised by different words and that their inclusion in a corpus to make a broad-coverage corpus is essential for an accurate representation of linguistic diversity in corpora. A broad-coverage corpus is a source of diverse linguistic wealth for dictionary compilation.

3.10 Concordance

Another way of studying words in a corpus is by studying a specific word in context in some detail in terms of co-texts to the left and to its right. This is achieved by generating a key word in context (KWIC) often referred to as concordance lines. “A concordance is an index of the surface word forms in a text. It is a collection of the occurrences of a word form, each in its own textual environment” (Dash and Chaudhuri, 2000: 190). A concordance reveals the company kept by a word, its collocates, and thereby reveal meanings and usages which are hard to dig up through mental recall. We illustrate this below through the example of the word *pelo* (heart*).*

Figure 1: Concordance results of the word *pelo*

```
o ka kgopolo ya gore Morwadi o tlaa wela pelo. A mo gaupanya. ka legofl fa ga re
 ngwatiaka, O se tshoge bono wa ka, Wela pelo ga o seitaodi re Use rotlhe, O sek
gang, o tla e rola morago o sena go wela pelo. '/'r~ a emelela, o b~ua a le esi)
a ka seatla. "0 sale sentle, moratiwa wa pelo ya me. Ga ke itse gore ke~ tla go
se o lela jalo? Ke a go rata moratiwa wa pelo ya me." Fa a sa ntse a e phimola,
ne a ithuta ona. "Gomotsega, moratiwa wa pelo, ya me." Mosele a didimala, mme go
 rebe la Mokwena, a buledisa moratiwa wa pelo ya gagwe. O ne a tsamaya ka bonya,
a. : Ke go reile ka re o seka wa utlwisa pelo botlhoko tlhe rra! 0 a itse gore b
```

```
g mo matlhong a gago. O se ka wa utlwisa pelo ya gago botlhoko ka nna, ke swetse
 gatlhisa thata., Ba,utlwil ba mo tswela pelo tota.. ,I, Mmaago Molebi a mo roma
be, a di phailela kwa, a re ba mo tswela pelo. Le ene Pule tota tsala ya gagwe y
hegelwa ke moratiwa, e seng go mo tswela pelo kgotsa go mo tlhoafalela. Seno se
sadi yo montle, mme phokojwe a mo tswela pelo. Phokojwe a leka maano a le mantis
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
ro ya gagwe. NtsVwa Mosetsanyana ya tswa pelo, ya metsa mathe a keletso. Saitsan
swe ke marago a tanka e nngwe. A tshwara pelo monnamogolo wa batho mme a ntse a
gotla-tshekelo. Mmaagwe Sereri a tshwara pelo ya gagwe, a bua ka tidimalo le bad
e motlha mongwe lokwalo lo tla tshikinya pelo ya ga Mmatheebe. A kwalela kgarebe
ologeletsweng morwadi wa we la tshikinya pelo ya gagwe gore a bale ka mabogo a a
nala kwa  Naledi a tla a ratile go tlola pelo. Mogoma e re ntlhomane a feta fa M
o ngwega Uncle Boot 0 ne a rata go tlola pelo. Mmaagwe ene, a ipega fa a ne a le
r ." Bikibiki a re, "Ngwana yo o tlhomola pelo. Lefatshe le mo itaya ka ntlha ya
 re ruri rre Rapitso wa batho o tlhomola pelo ka tshenyo e e leng ka mo supamake
 le bobotlana. Motho wa tsona o tlhomola pelo. Keikepetse o ntse jalo mo teramen
ka go tena Ontefile. Ketshedile a tlhapa pelo ka o ne a sa ntse a senka leano la
a. Bofelo a mo tsepela leitlho. A tlhapa pelo gonne fa go rata Bofelo, tsotlhe d
ng. jaanong Morwadi a nametsega a tlhapa pelo. A tsaya tlhogo a e latsa mo sehub
otlhapelo a ngwana wa mpa. "Nnake, tiisa pelo. O sa ntse o na le mogomotsi e bon
 bosigo. Tlogela go nna legatlapa. Tiisa pelo. Ga o sa tlhola o le mosimane. Ka
a gagamatsa thamo ya gagwe, a thatafatsa pelo ya gagwe, mo a bileng a se ka a bo
2Mme le ka sebaka seo Farao a thatafatsa pelo ya gagwe gape, a se ka a naya mora
wa gee." "Ke mang?" A botsa a swegaswega pelo. Ngwananyana a bolela fa ene a bon
a ke matlhagatlhaga, e bile a swegaswega pelo. O ne a batla go balela kwa pele.
```

The word *pelo*'s English equivalent is *heart* "a hollow muscular organ that pumps the blood through the circulatory system by rhythmic contraction and dilation" (Pearsall, 1998: 847). In the concordance lines above, *pelo* taken together with its collocates, is rarely used to convey the meaning of the physical heart. In the first line, *wela pelo* literally means "have your heart fall down" meaning "be at peace or settled." *Moratiwa wa pelo* (the loved one of the heart) is equivalent to "sweet heart" or "beloved". *Tshwara pelo* (handle or hold the heart) means "be in control of your emotions." It is through inspecting collocates that we can uncover proverbs, compounds, idioms, sayings, phrasal verbs and different multi-word expressions. Such structures could then be entered in dictionaries as sub-entries. Through the use of computer programs or concordance software it is relatively easy to get a list of all the cooccurences of a particular word in context and see all the meanings associated with the word (Biber et. al., 1998: 27). The concordance lines above reveal the different subtle meanings associated with the word *pelo*. From such a study of concordance lines, we have extracted possible subentries of the *pelo* headword. We have been able to extract 84 possible sub-entries (see Appendix 1); below we give only 10 of these.

Table 11: Corpus derived possible subentries of *pelo* entry

Collocates	Literal translation	Meaning
Ama pelo	Touch the heart	Hurt someone
Balabala ka pelo	Speak too much by heart	Talk aloud to yourself; absent minded
Baya pelo	Put the heart	Relax
Beta pelo	Suffocate the heart	Persevere
Betwa ke pelo	Be choked by the heart	Be very angry
Bofa pelo	Tie the heart	Restrain yourself
Bolawa ke pelo	Be killed by the heart	Desiring something
Bolwetse jwa pelo	The disease of the heart	Heart attack
Bona pelo	See the heart	See one's intentions or their thoughts
Bua ka pelo	Speak with the heart	To be troubled to the extent that you speak to yourself

The phenomenon of idiomaticity when considering a word and its collocates is not unique to the word *pelo* in Setswana. Words like *molomo* (mouth), *nko* (nose), *monwana* (finger), *kgomo* (cow) and *mpa* (stomach) all display similar characteristics. Such idiomatic expressions can enrich dictionary entries as subentries.

Table 12: Corpus derived possible subentries of *mpa* entry

Collocates	Literal translation	Meaning
Bana ba mpa	Children of a stomach	Relatives
Bipa mpa ka mabele	Cover the stomach with breasts	withhold bad information to protect a relative or friend
Gare ga mpa ya bosigo	In the centre of the belly of the night	In the middle of the night
Gare ga mpa ya lefatshe	In the centre of the stomach of the world	In the middle of nowhere
Gare ga mpa ya naga	In the centre of the belly of the wilderness	In the middle of nowhere
Mpa ya sebete	The belly of the liver	Flat on the stomach
Mpa e tuka molelo	A belly burning fire	Filled stomach
Go ja ka mpa tsoopedi	To eat with two stomachs	To eat until the stomach is full
Ntsha mpa	Take out a stomach	Commit abortion
Imelwa ke mpa	Be overladen with a belly	Have a full stomach

Table 13: Corpus derived possible subentries of *molomo* entry

Collocates	Literal translation	Meaning
Bolwetsi jwa tlhako le molomo	The disease of hoof and mouth	Foot and mouth disease
Itoma molomo wa tlase	Bite the lower mouth	Be determined
Itshwara molomo	Hold/touch a mouth	Be shocked
Ntsha ka molomo	Release with the moth	Speak

Pula molomo	That which opens the mouth	Money paid before someone speaks in lobola negotiations
Pipa-molomo	That which covers the mouth	A bribe
Rwala molomo	Carry the mouth on your head	To be angry and tight lipped
Roka molomo	Sew the mouth	Remain quiet
Tswa molomo	Grow mouth	Speak
Tlhoka molomo	Lack a mouth	Have nothing to say

Table 14: Corpus derived possible subentries of *lonao/dinao* entry

Collocates	**Literal translation**	**Meaning**
Apaya ka lonao	Cook with a foot	Avoid cooking and instead eat in other people's homes
Goga dinao	Drag feet	Move slowly
Fodisa dinao	Cool feet	Have a rest
Motsamaya ka dinao	One who walks with feet	A pedestrian
Ngotla dinao	Reduce feet	Reduce walking pace
Tlhatlosa dinao	Raise feet	Increase walking pace
Baya lonao	Put a foot	Be in a place
Tsholetsa dinao	Lift feet	Increase walking pace
Kgwele ya dinao	A ball of feet	Football
Tsosa dinao	Wake up feet	Increase walking pace/hurry up
Tiisa dinao	Strengthen feet	Increase walking pace

Table 15: Corpus derived possible subentries of matlho entry

Collocates	**Literal translation**	**Meaning**
Bula matlho	Open eyes	Educate/make aware/open eyes
Diga matlho	Drop eyes	Look down
Digalase tsa matlho	Glasses of the eyes	Spectacles/sunglasses
Latlhela matlho	Throw eyes	Look briefly
Matlho a phage a lebane	The eyes of a wild cat face to face	Face to face
Kala matlho	Measure eyes	Confuse
Tlodisa matlho	Make eyes jump	Overlook someone or something
Kgarakgaratsha matlho	Make eyes move from one place to another	Look from one place to another
Tlhatlosa matlho	Raise eyes	Look up
Tlhaetsa matlho	Shorten eyes from	Despise someone

Setswana dictionaries have attempted to include subentries based on the idiomaticity of collocates. However some of these have been few because of a lack of sufficient corpus evidence. Below we give examples of the treatment of *molomo* in different Setswana dictionaries.

Brown (1925: 210)

> **Molomo**, n., pl. melomo, A mouth (outside); a beak of a bird; a foreskin. *Kgwedi ea molomo*, the first month of the Sechuana year; the month of eating fruits. *Go cwa molomo*, to open the mouth, in speaking.

Kgasa (1976: 71)

> **molomo(me)** kgôrô e dijô di yang mo 'ganong ka yônê.

Kgasa and Tsonope's (1998: 171).

> **mo•lomo** TTT ln./3. me-. phatlha e e tswalwang ke dipounama tse pedi e go tsenngwang dijô ka yônê go ya ko mpeng le go bua. ♠ *molomo o tlola noka e tletse = motho o kgôna go bua dilô tse di ntsi tse a ka di dirang mme ntswa a se ka ke a kgôna*

Matumo (1993: 260)

> **molomo**, N. CL, 3 *mo-*. SING. OF *melomo*, a mouth; lip; a beak of a bird; an opening, as a tube, piping or tunnel; a foreskin. ID. EXPR., *go tswa molomo,* to open the mouth in speaking. PROV., *sejô sennye ga se fete molomo.*

Snyman et al (1990) does not enter *molomo*.

All the dictionary treatments of the *molomo* entry above are deficient and will benefit tremendously from the use of corpus evidence. For instance, the Matumo (1993) definition may be revised in the following way:

> **molomo**, *n.* 1. mouth 2. a lip. 3. a beak. 4. an object opening, as that of a bottle. ◘ **bolwetsi jwa tlhako le molomo**: foot and mouth disease. ◘ **itoma molomo wa tlase**: be determined. ◘ **itshwara molomo:** be shocked. ◘ **ntsha ka molomo**: speak; express an opinion; express a view. ◘ **pula molomo**: money paid before someone speaks in lobola negotiations. ◘ **pipa molomo**: a bribe. ◘ **rwala molomo**: be angry and tight lipped. ◘ **roka molomo**: remain quiet. ◘ **tswa molomo**: speak; say something; contribute; express an opinion. ◘ **tlhoka molomo**: Have nothing to say; be dumbstruck; be rendered speechless. ◘ **molomo o tlola noka e tletse**: it is easy for someone claim that they can achieve what they cannot do.

In the revised entry above ◘ is used to mark a subentry. Thus the study of collocations can enrich the dictionary entries. Thus we conclude this section by illustrating how dictionary entries for *pelo, mpa, matlho, molomo* and *lonao* could be enriched on the

basis of information in Tables 11, 12, 13, 14 and 15 derived from a corpus. We compare the proposed entries with entries from Matumo (1993). Matumo (1993: 306/7) enters twenty subentries for *pelo*. We have shown that over eighty sub-entries could be extracted from a corpus (see Appendix 1).

Matumo (1993: 276)

> **mpa** N. CL, 9Ø-, SING. OF *dimpa*, a belly; a stomach. ID. EXPR. *mpa ya lentswê*, the middle of a hill; *mpa ya lonao*, the sole of a foot. PROV., *seboba re bata sa mokwatla sa mpa re a mpampetsa.*

Matumo's *mpa* entry might be improved in this way:

> **mpa** *n.* a belly; a stomach. ◘ **bana ba mpa**: relatives ◘ **bipa mpa ka mabele**: withhold bad information to protect a relative or friend ◘ **gare ga mpa ya bosigo**: in the middle of the night ◘ **gare ga mpa ya lefatshe/naga**: In the middle of nowhere ◘ **mpa ya sebete**: flat on the stomach ◘ **mpa e tuka molelo**: with a full stomach ◘ **go ja ka mpa tsoopedi**: to eat until the stomach is full ◘ **ntsha (senya) mpa**: commit abortion ◘ **imelwa ke mpa**: have a full stomach.

Matumo (1993: 212)

> **lonaô** N. CL. 11 *lo-*, SING OF *dinaô*, a foot. ID EXPR, *go baba lonaô*.

Matumo's *lonao* entry might be improved in this way:

> **lonao** *n.* a foot ◘ **apaya ka lonao**: avoid cooking and instead eat in other people's homes ◘ **goga dinao**: move slowly ◘ **fodisa dinao**: have a rest ◘ **motsamaya ka dinao**: a pedestrian ◘ **ngotla dinao**: reduce walking pace ◘ **tlhatlosa dinao**: increase walking pace ◘ **baya lonao**: be in a place ◘ **tsholetsa dinao**: increase walking pace ◘ **kgwele ya dinao**: football ◘ **tsosa dinao**: increase walking pace ◘ **tiisa dinao**: increase walking pace.

Matumo (1993: 232)

> **matlhô** N. CL. 6 *ma-*, PL OF CL. *leitlhô*; *maitlhô* is still used in a few areas, eyes.

Matumo *matlho* entry might be improved in this way:

> **matlhô** *n.* eyes. ◘ **bula matlhô**: educate, make aware, enlighten ◘ **diga matlhô***:* look down ◘ **digalase tsa matlhô**: spectacles, sunglasses ◘ **latlhêla matlhô**: look briefly ◘ **matlhô a phagê a lebane**: face to face ◘ **kala matlhô**: confuse ◘ **tlodisa matlhô**: overlook someone or something ◘ **kgarakgaratsha matlhô**: look from one place to another ◘ **tlhatlosa matlhô**: look up.

By proposing improvements for the dictionary entries, we hope to have illustrated the power of corpus evidence and concordance lines. However corpus generated collocations and frequency lists have not always been used to inform the complexity of a dictionary entry. Other methods have been explored which we discuss briefly below.

3.11 A review of existing methods of headword list identification

We have argued how a corpus can be used as a source of headword and subentries. However, a corpus has not and is not always used by lexicographers for dictionary compilation. In this section we review different methods which have been used by lexicographers to identify headword lists for dictionary compilation. For ages, lexicographers battled with ways and means of producing authentic and reliable reflections of the lexicon for different languages. Most of these lexicographers depended on their ability to remember words that existed in the languages under study, something that Prinsloo and De Schryver (2000: 4) call entering "words as they cross the compiler's way" and Kilgarriff (2000: 109) call "the lexicographer's intuition". Others on the other hand, in the Oxford tradition, depended on readers, who searched texts for word occurrences and submitted citations of words for entry into the dictionary. The readers' contribution, for many years, made the OED (Oxford English Dictionary) the most comprehensive lexicographic work of the English language. Developments in lexicography, later proved that readers were not reliable sources of dictionary material since they did not only take too long to process data, but they also could not accurately deliver information on matters of frequency across texts and genres to aid decisions on what to include and exclude (Summers, 1995, Sinclair, 1996 and Kilgarriff, 1997: 135).

Since the revolutionary COBUILD research using corpora evidence of 1981 (Sinclair, 1987 and Sinclair, 1991 and Moon, 2007: 159) there has been a rapid increase of dictionary projects that depend on corpora. The earlier Birmingham school of corpus lexicography adhered religiously to a corpus as a source of dictionary evidence. They argued that corpora were the sole sources of lemmatisation, frequency information wordlists and authentic examples (Fox, 1987: 138/9). If a word was not in a corpus it was not recognised as legitimate dictionary material.

However as corpus lexicography develops, the focus does not lie on corpus output exclusively, but more crucially, on corpus design and composition since they determine corpus output. Matters of representativeness, balance and genre coverage become more urgent to both theoretical and practical lexicographers (For a detailed discussion of this matter, refer to Chapter 4 of this thesis). Researchers want to know the nature of texts that form a corpus and in what proportions they stand to each other (Kilgarriff, 1996). Therefore the greatest challenge lies not so much in what we get from a corpus, but rather in its construction, for it is what goes into a corpus that determines what can be extracted from it.

For the remainder of this chapter we focus the discussion on the headword list identification. We begin by sketching the English tradition of headword list identification. We then proceed to discuss the non-corpus approaches to dictionary compilation and end by looking at corpora use in Setswana dictionary compilation.

3.12 A historical perspective of headword lists

The earlier English dictionary compilations where characterised by two phases; the Latin-English dictionaries and the dictionaries compiled by direct borrowing from literary works, especially technical terms appended to learned vernacular publications of the time.

The need to list words may be traced to the seventh and eighth centuries when "priests and scholars, glossing Latin manuscripts, compiled lists of difficult words to help readers unfamiliar with Latin" (Wells, 1973: 13). These lists grew longer and were subsequently presented in alphabetical order for easy access. They developed into what became Latin-English, English-Latin bilingual dictionaries. This laid the foundation for what could be termed dictionaries of "hard words" tradition of the 16th century.

For African languages it was the European explorers (Naden, 1993; Lichtenstein, 1928-30) and missionaries (Moffat, 1826) in the 1800s who recorded languages either out of curiosity or for Bible translation purposes.

By the 17th Century, in the English tradition, it was clear that lexicographers depended on reading many books for compiling wordlists as Bailey (1736) states in his preface that he depended on "the reading of a very large number of authors…" (quoted in Wells, 1973: 21). This approach was also adopted and developed by Samuel Johnson who "added a new empiricism, a wide ranging program of reading diverse sources" (op. cit. 21). Bailey compiled a 40,000 entry dictionary, the *Universal Etymological English Dictionary* (1721) (see Osselton, 1983).

The problem of collecting words has posed great difficulties to English lexicographers for a long time, as it currently does to Setswana lexicographers, especially for those who were interested in not merely copying other dictionaries.

> But to *collect* the *words* of our language was a task of greater difficulty: The efficiency of dictionaries was immediately apparent; and when they were exhausted, what was yet wanting must be sought by fortuitous and unguided excursions into books, and gleaned as industry should, or chance should offer it, in the boundless chaos of a living speech. My search, however, has been either skilful or lucky; for I have much augmented the vocabulary (Johnson, 1963: 10).

Johnson here points to three sources he used for his dictionary: other dictionaries, books, and "living speech". Some of the texts were from as diverse sources as science, technical dictionaries and philosophical writings.

> On the terms of art I have received as could be found either in books of science or technical dictionaries; and have often inserted, from philosophical writers, words which are supported perhaps only by a single authority, and which being not admitted into general use, stand yet as candidates or probationers, and must depend for their adoption on the suffrage of futurity (op. cit.).

Although Johnson's method was not completely corpus-based, it does point to a dependence on diverse sources of texts and spoken language for dictionary material.

3.13 Non-corpus dependant methods of dictionary compilation

Although in this thesis we argue for corpus methodology in dictionary compilation, there are still many practising lexicographers who use other strategies in compiling dictionaries. Ooi (1998: 47/48) identifies two different ways in which lexical or lexicographic evidence is derived for inclusion in dictionaries. These are lexical introspection and casual citation. By lexical introspection is meant the lexicographer's linguistic introspection, the words he can remember. Casual citation refers to "when the lexical behaviour of one's family members, friends, or strangers is observed and recorded" (op. cit. 48). In this instance lexicographic evidence is based on the people a lexicographer comes into contact with.

Other dictionaries have been used as sources of lexicographic evidence as seen in Section 3.12. Zgusta argues that "An important source of information can be found in other dictionaries of the language in question, if there are any" (Zgusta, 1971: 239). For the compilation of old dictionaries and even in some modern dictionary compilation practices, lexicographers have copied other dictionaries. The practice predates Johnson's dictionary. Thus Wells notes that,

> Lexicographers have traditionally borrowed quite freely from preceding dictionaries, sometimes plagiarizing with a free hand… [and] more often existing dictionaries were consulted and synthesized with other sources such as spelling books and technical glossaries (Wells, 1973: 21).

However Svensén warns against this practice noting:

> …there is a type of evidence which can be misleading, and curious enough is to be found in dictionaries. Straightforward errors in one dictionary may, when this in turn is used as a source for another dictionary, come to be regarded as authentic linguistic productions. This gives rise to 'ghost words', i.e. words which do not actually exist (Svensén, 1993: 41).

But Svensén's caution is not new; Johnson saw the weaknesses of his predecessors in including words which could not be accounted for anywhere in written texts or speech, and decided to omit them from his dictionary. Bailey, Ainsworth and Philips are lexicographers who had published dictionaries before Johnson.

> Many words yet stand supported only by the name of Bailey, Ainsworth, Philips, or the contracted *Dict.* for *Dictionaries* subjoined: of these I am not always certain that they are read in any book but the works of lexicographers. Of such I have omitted… (Johnson, 1963: 12/13).

Another method for gathering text dictionary compilation is the use of semantic domains developed by Ronald Moe of the SIL (Summer Institute of Linguistics).

3.14 Semantic domains

Moe (2001) of SIL (Summer Institute of Linguistics) proposes a method of semantic domains to be used for the collection of words. He argues that the methodology is particularly attractive for minority languages, most of which have none or few written texts, or no corpora. His argument is that the methodology is 100 times faster than collecting words without a structure. He argues that 12,000 words have been collected in a few weeks through what is effectively a simple methodology but one which is able to produce a massive classified dictionary and thesaurus.

Moe analysed domain classification of words as suggested by Murdock et al. (1987), Roget's (1958 and 1985 editions) and Louw and Nida (1989) and found them inadequate for eliciting vocabulary. What Moe attempts to compile is "a universal list of semantic domains" (Moe, 2001: 151) which field lexicographers could use to prompt native speakers to think of words in their language. However, semantic domains have greater relevance than mere elicitation of mother tongue speakers' words. "It could be used to collect words, it could serve to classify a dictionary, and it could aid in semantic investigation" (Moe, 2001: 152). Underlying this system is a mental approach to the lexicon; that words are all linked together in the mind in a gigantic multi-dimensional web of relationships which cluster around a central nexus (Moe, 2001: 4). The mental

lexicon is not alphabetical but words cluster around key concepts and it is these concepts that Moe calls semantic domains (Moe, 2003: 216). It is therefore his argument that related words should be collected at the same time. To guide field workers, Moe phrases domains as questions as in the following for the domain 'sing":

What words refer to singing? sing, serenade, warble, yodel, burst into song
What words refer to singing without using words? hum, whistle

These series of questions are central to what Moe calls the Dictionary Development Process (DDP) which he used in Uganda in training lexicographers in collecting Lunyole (a Bantu language) words. The DDP has 1,700 domains each with 8-10 questions which could elicit over 10 words per domain which means that the dictionary would have at least 17,000 entries.

Moe's semantic domain approach is relevant for the kind of context for which it has been constructed – minority languages with a very limited written tradition. It will prove very useful for individuals who gather words in rural areas and communities with languages with none or limited written tradition, where there may be a lack of written texts and technology to capture and process oral data of such languages. The semantic domain approach may also be used to augment a wordlist compiled from what could be perceived as an imbalanced corpus as a result of a lack of text from a specific genre.

While the semantic domain method may be used for gathering words of lesser-known languages of the world with limited or no written tradition, for languages with a large body of written texts this tedious task may not prove essential since huge corpora could be compiled which could be queried cheaply in various sophisticated ways. The semantic domain method of lexical collection does not provide any frequency information. Rather it in effect enters words into a dictionary as they are remembered by respondents.

For the purposes of this thesis we favour data from a well designed corpus. By a well designed corpus we refer to a corpus that comprises samples of language varieties from a language of interest. Central to the use of corpora is that linguistic information that goes into making a dictionary 'must be authentic, that is to say it must include only such

linguistic occurrences as actually exist… the lexicographer must find evidence for it in independent sources" (Svensén, 1993: 40).

3.15 Corpus lexicography and Setswana dictionaries

Of the entire Setswana dictionary compilations discussed in Chapter 2 of this thesis, it is Kgasa and Tsonope (1995) who report the use of a corpus in the compilation of their dictionary. They point out that:

> Re dirisitse tsa maranyane a dikhompiutara go tlhotlha le go runa mafoko a feta dikete di le makgolo a mabedi le masome a matlhano (250, 000) mo dikwalong di le mmalwa; ra tloga ra a oketsa ka mafoko a mediriso-puo e e faphegileng jaaka maina a dinaledi, dinonyane, mebala le matshwao a diphologolo, ditlhare, ditlhaga le dimela tse dingwe, ditiro tsa Setswana jalojalo (Kgasa and Tsonope 1995: v-vi).

> We have used computer technology to analyse a corpus of more than 250, 000 words from a compilation of several books; we then added special terms such as names of stars, birds, animal colour terms, names of trees, grasses and other plants, and terms particular to the Setswana culture etc (translation mine).

Kgasa and Tsonope are pioneers of corpus use in Setswana lexicography, particularly in Botswana. Their compilation of a corpus of a quarter of a million Setswana tokens in 1990 was an enormous achievement in an environment where Setswana language texts were not readily available.

While their corpus compilation is commendable, little is known about its structure and quality since their corpus construction process is not documented in any publication that we are aware of save for what we have quoted above from the introduction to the dictionary.

It is also clear that Kgasa and Tsonope did not sample any spoken language for their corpus. This is not surprising since the compilation of spoken corpora is both tedious and expensive. However a lack of transcribed speech in corpora has lead to deficiencies which have been observed in the literature. We had argued against such an imbalance:

> Such an imbalance raises questions relating to the composition and balance of the corpus. This is so since speech is the primary channel of human communication and exists in abundance compared to written text. (Otlogetswe, 2004: 194).

We have also argued (Otlogetswe, 2006: 150-153) that the exclusion of spoken text results in a loss of instances of borrowings from Afrikaans and English which are not usually accepted in the written Setswana form by many publishers.

Other dictionaries like *Dikišinare ya Setswana English Afrikaans Dictionary Woordeboek* are purely introspective in their approach. No wonder the compilers say:

> The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compiling of a dictionary. This can take place simply because the lexicographer had not encountered such words. We can only hope that there are not too many examples of this kind (Snyman et al., 1990: preface).

Matumo's (1998) is an updated version of Brown (1925). It does not make any claim of corpus use.

3.16 Conclusion

In this chapter we have demonstrated how frequency lists aid in the determination of words commonly used in a corpus. We have argued that such information can assist the lexicographer to compile headword lists. We have illustrated how the exclusion of functional words from the frequency list may reveal clearly the words that are typical of a corpus or subcorpus. While frequency lists have proved to be useful in the identification

of genre-specific words they have been found limited. We have shown how keyword analysis assists in the isolation of genre specific words which could form part of a headword list. Keyword analysis is also significant in genre identification in sociolinguistics and lexicography. We use Keyword analysis in detail in Chapter 6. In addition to frequency analysis, we have argued that a corpus can be exploited through concordance lines. Concordance lines are significant in revealing words which occur in the company of others. They unearth collocation, idiomatic expressions, phrasal verbs and various multi-word expressions.

Chapter 4

Issues in corpus design for lexicography

> One of the main issues addressed here, though, is whether general language studies must be based on a corpus that is register-diversified as well as large (Biber, 1993: 220).

4.1 Introduction

In the previous chapter we have considered what a corpus is and a variety of ways in which it is exploited for different ends. In this chapter we look at issues which arise in corpus design, particularly as they relate to the area of lexicography. Corpus design is relevant to this thesis since at the heart of this thesis is the argument that corpus design and compilation determine the quality of what could be extracted from it. The area of corpus design is broad and an attempt will be made to cover some of its most fundamental matters. Atkins et al. (1992) present a detailed discussion on corpus design through a panoramic overview of corpus design including practical stages of compiling a corpus including text selection and mark-up; the problems of defining a population of texts to be sampled; the types of corpora and their various uses. Some of the issues they raise will be investigated in considerable detail in this chapter.

As the use of computer-based text corpora has become increasingly important for research in natural language processing, lexicography, and descriptive linguistics, issues relating to corpus design have also assumed central importance (Biber, 1993: 219). Therefore a "corpus which is designed to constitute a representative sample of a defined language type" (Atkins et al., 1992: 2) has become increasingly attractive. Samples may be divided into two broad categories of written and spoken text. Written text refers to such written products as books, novels, magazines and letters. Spoken text refers to transcribed speech from meetings, lectures, telephone conversation, interviews or

debates. These two broad categories are characterized by variability.

It is a linguistic truism that language is characterized by varieties (Fromkin and Rodman, 1998: 400-404). These varieties may be sociolects or social dialects, that is, linguistic varieties on the basis of facts such as socioeconomic status, gender, ethnic grouping, age, occupation and others (Southerland and Katamba, 1996: 540). There are also regional varieties; distinct linguistic varieties which characterise people from a certain geographic area. Linguistic varieties may also be perceived from the perspective of functional speech varieties also known as registers which characterise language on the basis of whether it is casual, formal, technical and other characteristics (Hudson, 2000: 452).

The recognition of a lack of linguistic uniformity in speech communities has relevance to corpus design since it means that "…due to the importance and systematicity of the linguistic differences among registers, diversified corpora representing a broad range of register variation are required as the basis for general language studies" (Biber, 1993: 219). We therefore differ with some proponents of very large corpora who have "suggested that size can compensate for a lack of diversity – that if a corpus is large enough, it will represent the range of linguistic patterns in a language, even though it represents only a restricted range of registers" (Biber, 1993: 220).

The design of corpora for lexicography comprising a diversity of texts raises multiple issues which are the subject of this chapter. These matters include amongst others: balance and representativeness, corpus size, corpus annotation, sample size and spoken language in a corpus. We begin by the subject of balance and representativeness.

4.2 Balance and representativeness

Biber (1995: 130/131) notes that in the area of social sciences, issues of representativeness are dealt with under the rubric of *external validity*, which refers to the extent to which it is possible to generalize from a sample to a larger target population. However there are two kinds of error that can threaten external validity: *random error* and *bias error*. *Random error* occurs when the sample is not large enough to accurately

estimate the true population; *bias error* occurs when the selection of a sample is systematically different from the target population. Random error can be minimised by increasing the sample size, and this is why large text corpora are important. Bias error on the other hand refers to the sampling of only a part of a population to the exclusion or limited inclusion of other parts of the population. In contrast, bias error *cannot* be reduced by increasing the sample size, because it reflects systematic restrictions in selection. That is, regardless of corpus size, a corpus that is systematically selected from a single register or limited varieties cannot be taken to represent the patterns of variation in an entire population. Rather, in order to make global generalizations about variation in a language, corpora representing the full range of registers are required. Bias error therefore has to be addressed by broadening the representation of linguistic variability in a corpus.

The matter of balance and representativeness is one of the greatest areas of contestation in corpus design and compilation. On one hand, there are those who argue that a language can be sampled in its varieties to form a corpus that can be taken as a representative sample of the whole language. For instance, Renouf points out that:

> When constructing a text corpus, one seeks to make selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested (Renouf, 1987: 2).

There are those who argue that since we can never know all the varieties of a language and researchers possess no facts about the amount of spoken or written text that exist in real life, there is no way anyone can claim to compile a corpus that can be representative of the whole language. We explore both arguments.

4.2.1 Proponents of balance and representativeness

Biber et. al. (1998: 246) state that a corpus is not just a collection of texts, but at the heart of corpus design and construction is an attempt at creating a representative sample of a language or parts of a language that can be studied. Representativeness here

according to Biber should be understood to mean "the extent to which a sample includes the full range of variability in a population" (Biber, 1994: 378). The "full range of variability" here refers to the range of text types and of linguistic distributions in a language. Therefore this means the object that is represented needs to be well understood by a compiler since "an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from the population" (Biber, 1994: 378). This position is similar to the one held by Renouf (1987: 2) who argues that "The first step towards this aim [constructing a corpus] is to define the whole of which the corpus is to be a sample." Biber et al. show that one of the problems in sampling is characterising the language to be sampled. However one of the limitations of attempting to characterise the language is that "we do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all variation in texts" (Biber et al., 1998: 246).

While the full varieties of a language may be unknown, there are other simpler cases where the whole text to be analysed may be finite and known as in the case of the total works of Shakespeare or the whole Bible text (Renouf, 1987: 2). Kilgarriff and Grefenstette however contend that, "A corpus comprising the complete published works of Jane Austen is not a sample, nor is it representative of anything else" (Kilgarriff and Grefenstette, 2003: 334) since it is the complete works of a specific writer.

Language can also be sampled proportionally. Such sampling will translate to highly used varieties sampled in greater proportions compared to rarely occurring ones. This will mean that since speech is used more in human communication compared to written language, corpora would have higher levels of spoken language compared to written language. A corpus designed in this manner approximates Biber's rough estimates:

> A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing (Biber, 1994: 386).

Such a corpus could be considered representative only in that it approximates how

different varieties are used in a language. Biber (1994) however argues that proportional representativeness is not interesting for linguistic research. What is interesting however is "language samples that are representative in the sense that they include the full range of linguistic variation existing in a language." The major weakness with proportional sampling of language (i.e. both produced and received language), Biber has argued, is that even if it could be achieved, it would result with relatively homogenous corpora. This is because most texts in such corpora would be from conversation therefore having similar linguistic characteristics, since speech is proportionally greater than written language. A proportional sample may therefore not include texts from registers which are rarely read by the public such as legal and medical documents (cf. Burnard, 1995).

Biber et al. (1998: 89) therefore point out that a "key aspect of corpus design for most studies, then is including the range of linguistic variation that exists in a language, not the proportions of variations." They argue for *stratified* sampling which involves cataloguing the different categories of texts that exist in a language and sampling each of them, instead of *proportional* sampling which tries to compile proportions of language varieties that people use and receive.

Their argument is therefore that corpus language variability must approximate the linguistic variability of a speech community under study or if it does not, corpus limitations should be acknowledged. Biber (1995: 27) notes that in the sampling of a language,

1. the full range of registers in the language should be included, representing the range of situational variation
2. a representative sampling of texts from each register should be included; and
3. a wide range of linguistic features should be analysed in each text, representing multiple underlying parameters or variation.

Here Biber argues for the representation in a corpus of the intricate varieties of a language under study, for if a corpus lacks the major text types, genres or dialectal varieties, it cannot be said to represent the general language. Furthermore, Leech argues that:

> The value of a corpus as a research tool cannot be measured of brute size. The **diversity** of the corpus, in terms of the variety of registers or text types it represents, can be an equally important (or even more important) criterion. So, too, can the care with which it has been compiled…" (Leech, 1997: 2, emphasis in the original).

Register diversity is therefore crucial in a corpus to ensure the faithful representation of linguistic variability found in a language.

While Biber et al. argue against proportional representation, Rayson (2002: 42) contends that for a corpus to be representative of the language as a whole, it should contain samples of all major text types and, "if possible, be in some way proportional to their usage in every day language." This sense of representativeness is different to that suggested by Biber (1994 and 1998) since while he argues for the inclusion of the diversity of text types in a corpus; Rayson argues that such samples should be in some way proportional to the varieties used in a language.

Corpus linguists and corpus lexicographers consistently argue for representativeness in corpus construction mainly because for corpus results to be generalized to the whole language, the corpus must be seen to be compiled in a systematic manner that is perceived to be representative of the population from which it was abstracted to justify the generalizations. Summers points to the functionality of corpus representativeness when she says:

> One of the many reasons for wanting the corpus to be representative was so that reliable frequency statistics could be generated and used to aid the lexicographers in making the many linguistic judgements that lie behind the final entry for a word in the printed dictionary (Summers, 1996: 261).

The lexicographer's linguistic judgements aided by frequency statistics that Summers refers to, include amongst other things how to frame an entry, the ordering of definitions in the entries and the sub-entries of a headword (see Chapter 3, section 3.5). Such authoritative decisions may be reached through the exploitation of corpora.

Biber also expresses a similar position to that of Summers. He argues that "a corpus must be representative in order to be appropriately used as the basis for generalisations concerning a language as a whole" (Biber, 1993: 243).

It is clear that a representative and balanced corpus must represent the different genres of language use in a language community. According to those who argue for proportional sampling, a representative and balanced corpus would additionally attempt to capture the proportions, that is, different ratios of the different varieties in a specified language community. The determination of proportions is hard to achieve, as Biber (1998) has shown mainly because it is difficult to know precisely all the text types and their proportions of use in a population with its ever-changing dimensions. The difficulties are compounded when one faces the compilation of a corpus of spoken language. This is the case since as Kilgarriff (1997: 137) points out dialectal varieties stand at different ratios to one another and should be represented within a corpus that attempts to accurately capture the language dimensions as a whole.

4.2.2 A cautious approach to balance and representativeness

On the other hand, Kennedy is not convinced that the representativeness ideal can be achieved in a corpus.

> The extent to which a corpus can ever be considered to represent a language in general is currently a matter of some contention. In practice, whether a finite sample of a language could ever 'represent' the vast amount of a language produced in even a single day is always likely to be, in the final analysis an act of faith (Kennedy, 1998: 21).

Kennedy (ibid: 62) is additionally doubtful that we can confidently argue for representativeness of a corpus that represents a language.

> In light of the perspectives on variation offered by several decades of research in discourse analysis and sociolinguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even

of a particular genre or subject field or topic (Kennedy, 1998: 62).

By "perspectives on variation" Kennedy refers to different speech varieties that exist in a speech community. He is referring to challenges faced by sampling the standard against non-standard varieties; various sociolects covering socioeconomic status, gender, ethnicity, age, occupation, and others; different regional varieties, like *Sengwaketse*, *Sekgatla*, *Sekwena*, *Sengwato* in the case of the Setswana language; different registers like casual, formal technical and others. Such variations are difficult to represent in a corpus. By noting this difficulty, Kennedy does not imply that representativeness should not be attempted, but that perhaps theoretically an attempt at representativeness may not conclusively capture the nuances of existing varieties as perceived in linguistic research. He therefore concludes that "a 'representative' sample is at best a rough approximation to representativeness, given the vast universe of discourse" (Kennedy, 1998: 52).

Rundell also reveals the practical challenges of achieving representativeness and balance:

> In practice, it is not always feasible to assemble precisely the corpus one ideally wants: practical constraints, such as a shortage of time and money, the variable availability of machine-readable text, and problems with copyright clearance, all conspire to make compromises necessary (Rundell, 1996, online[7]).

It is precisely the problems outlined by Rundell, which stand out as some of the major impediments particularly in the African context to corpus construction. The lack of machine readable data, the unavailability of funding, the demanding transcription of spoken language and cleaning of scanned texts remain as hurdles to building corpora that capture linguistic variability of a specific linguistic community.

In compiling the BNC, Burnard notes that the objective was to define a stratified sample according to stated criteria, so that while no-one could reasonably claim that the corpus

[7] http://www.ruf.rice.edu/~barlow/futcrp.html

was statistically representative of the whole language in terms of either production or reception, at least the corpus would represent the degree of variability known to exist along certain specific dimensions, such as mode of production (speech or writing); medium (book, newspaper etc.); domain (imaginative)… (Burnard, 2002: 60).

Burnard emphasises the difficulty of attempting linguistic representativeness in a tight statistical sense, but rather that corpus representativeness for the BNC was determined in terms of known linguistic varieties, a position similar to the one held by Biber (1994).

A corpus intended to represent the "general language" but lacking in linguistic variability can lead to erroneous conclusions. Ooi argues that "a corpus selected wrongly or inadequately runs the risk of generating not only 'noise' in the information acquired but not offering any information at all" (Ooi, 1998: 52). Take for instance Verlinde and Selva (2001) who compare the corpus-based and intuition-based lexicography in French lexicography. They note that although the French lexicographers were some of the first to incorporate corpus approaches to dictionary making the lexicographic landscape in France has largely remained intuition-based. They use 50 million words of the 1998 issues of *Le Monde* and *Le Soir* to draw up a frequency list and make comparisons between the corpus list and dictionary entries. For their electronic French learner's dictionary they decided to limit the selection of their lemmas to 12 156 words by including only those lemmas that occurred at least 100 times in a 50 million-word corpus. Since it is a learners' dictionary certain words were excluded. Amongst these were words found in current affairs like *bosnique*, *kosovar* and *brainois*. By running frequency lists they identified that 12% of the 12,000 most frequent words of their corpus did not occur in *Dictionnaire du français*. They thus concluded:

> Corpus-based lexicography gives strong and necessary empirical evidence to the lexicographer's personal intuition, even if this personal intuition remains helpful in filling the gaps in our corpus (Verlinde and Selva, 2001: 598).

While they make a valid point concerning corpus-based lexicography, at least one point of criticism may be made in relation to Verlinde and Selva's experiment on the basis of the nature of the corpus they used.

Although they admit that central to corpus building are the matters of corpus representativeness and size, for them to "rely on the texts that are freely accessible" (Verlinde and Selva, 2001: 594) and in this case, text from two newspapers, defeats the point of representativeness that they attempt to defend. Biber arguing for his Multi-Dimensional (MD) approach to studying language variation has shown that a single register cannot be said to represent broad linguistic variability of a language.

> That is, regardless of the corpus size, a corpus that is systematically selected from a single register cannot be taken to represent the patterns of variation in a language, corpora representing the full range of registers are required. For MD analyses, it is important to design corpora that are representative with respect to both size and diversity. However, given limited resources for a project, *representation of diversity is more important for these purposes than representation of size* (Biber, 1995: 131, italics mine).

Biber's view equally applies to corpora designed for lexicography. An admission with qualification by Verlinde and Selva (2001: 594) that: "We cannot say that our corpus is perfectly balanced, but it is made up of the kind of texts that the potential users of our dictionary will have to deal with" undermines the linguistic variability found in different genre and text types since the 50 million-word corpus is highly skewed towards one kind of genre, namely, newspaper text. Their frequency lists are not compelling although extracted from a huge corpus. The corpus lacks texts from domains such as novels, magazines, radio interviews, textbooks, sports commentaries, film, poetry, speeches and spoken text, which we expect dictionary users to encounter daily. Since their corpus lacks text variability, their criticism of *Dictionnaire du français* that it lacks certain words found in their frequency list may only be because of the inadequacy of their corpus rather than the introspective lexical inclusion principle on the part of *Dictionnaire du français* compilers. Verlinde and Selva could have evaluated their list to ascertain that it captured words from cross the spectrum of French language use. Additionally, research needs to be conducted on the degree of linguistic variability in newspaper text compared to corpora compiled from a variety of text types.

Sinclair (2004) cautions against claims of mathematical exactness in language sampling by arguing that,

> We should avoid claims of scientific coverage of a population, of arithmetically reliable sampling, of methods that guarantee a representative corpus. The art or science of corpus building is just not at that stage yet, and young researchers are being encouraged to ask questions of corpora which are much too sophisticated for the data to support. It is better to be approximately right, than to be precisely wrong (Sinclair, 2004).

Sinclair's position does not mean that he opposes representative corpora or that corpora cannot be representative, for he argues that "The contents of the corpus should be chosen to support the purpose, and therefore in some sense represent the language from which they are chosen" (Sinclair, 2004). However what he opposes is the assumption that the population is well defined, fully known and perfectly understood.

Sinclair (2004) also argues that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within and outside it that cause it to develop continuously. As a consequence he contends that no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. This position is similar to that of Biber et al. and Kennedy discussed earlier who argue respectively:

> ….we do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all variation in texts (Biber et al., 1998: 246).

> and

> …. it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic (Kennedy, 1998: 62).

Sinclair therefore argues that corpora researchers sample, like all the other scholars who study unlimitable phenomena. He argues that:

> We remain, as they (scholars who study unlimitable phenomena) do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population (Sinclair, 2004: online[8]).

By arguing against proportional representation Sinclair agrees with Biber et al. (1994) who argue for stratified and non-proportional sampling.

He argues that to discuss the concept of representativeness we must consider the users of the language we wish to represent and ask ourselves the following questions:

- What sort of documents do they write and read, and what sort of spoken encounters do they have?
- How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications?
- How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web-pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence?
- How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

(Sinclair, 2004: online[9])

Such questions will guide a compiler in selecting relevant text to include in the corpus.

Sinclair (2004) again is helpful in pointing out that “The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.”

---

[8] http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

[9] http://www.ahds.ac.uk/creating/guides/linguisticcorpora/chapter1.htm

Kilgarriff and Grefenstette (2003: 340) echoing Kennedy (1998: 62) argue that "representativeness" begs the question "representative of what?" The problem of what is represented by corpora is particularly compounded by designs of corpora of "general language" which is hard to define. Representativeness therefore raises serious theoretical issues about language modelling including issues such as:

- *Production and reception*: is what is modelled received (read and heard) or produced (written and spoken) language or both? The British National Corpus, for instance, attempted to take care of both perspectives (Burnard, 2002: 22).
- *Balance between speech and text corpus amounts*: We must also contend with whether spoken text can be accurately sampled and represented along the same lines as written text. How many words are we looking for and what percentage of the spoken language do such words constitute? Whether spoken text can be sampled in any representative manner is greatly questionable. While we can sample *Sengwaketse*, *Selete, Sengwato, Sekwena,* or *Sekgatla* dialects in the Setswana language, establishing an acceptable representative percentage of the spoken form of these dialects poses great difficulties since as we attempt to quantify them, more speech instances are produced. Even if we settled for a stratified sampling, we are left with the question of, how much from each stratum?
- *What constitutes distinct language events*? Do repetitions, copying, quotation, or republications of similar stories in different newspapers constitute distinct language events that could be represented in a corpus?

With the haze that clouds matters of representativeness and balance, and with limited understanding of text types, genres language varieties in research, Kilgarriff and Grefenstette, writing about using Web text as corpus, argue that:

> The web is not representative of anything else. But nor are other corpora, in any well-understood sense. Picking away at the question exposes how primitive our understanding of the topic is, and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it may be modelled (Kilgarriff and Grefenstette, 2003: 343).

Kilgarriff and Grefenstette argue that corpora if well understood cannot be said to be

representative of anything else.

So far we have attempted to show the complexity of matters of balance and representativeness and how researchers differ on whether language can be sampled in a represented manner. As Sinclair (2004) has noted, one major complicating factor in building balanced and representative corpora is that language is an "unlimitable phenomena". It is unknown how many words or sentences exist in writing or how many have been uttered or will be uttered. A quest to quantify such data would result in general estimates, for more publications are produced every minute and speech is continuously produced. Such recognition of language as an unlimitable phenomenon however does not obstruct researchers from arguing for sampling different linguistic varieties for both quantitatively and qualitatively inspection. The challenge for corpus linguists and lexicographers is to identify the parameters of a language to be studied and sample them for corpus analysis. Sinclair (2004) suggests the following ways of achieving representativeness in a corpus:

1. decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components;
2. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;
3. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type;
4. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;
5. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;
6. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

(Sinclair, 2004: online[10])

[10] http://www.ahds.ac.uk/creating/guides/linguisticcorpora/chapter1.htm

While it may be difficult to define and accurately characterise balance and representativeness, most modern corpus based lexicography research still consider issues of representation and balance (Ooi, 1998) as marks of standards of authenticity and robustness in corpus construction as Sinclair shows:

> The notion of balance is even more vague than representativeness, but the word is frequently used, and clearly for many people it is meaningful and useful. Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements (Sinclair, 2004).

Reményi (2001: 486) argues that "the problems of 'representativeness' are mostly due to the double nature of the unit of observation in corpus design: either the diversity of language users, or that of text types is eclipsed." The problems lie in whether language users (text producers and receivers) or texts (the products of language use) be chosen as the units of observation. Additionally corpora organised by demographic proportions would not support the criterion of 'sample variability matching population variability' as far as text types are concerned.

Atkins et al., introduces the concept of *organic corpora*, as a possible approach of addressing matters of representativeness and balance.

> A corpus builder should first attempt to create a representative corpus. Then this corpus should be used and analysed and its strengths and weaknesses identified and reported. In the light of experience and feedback the corpus is enhanced by the addition or deletion of material and the circle repeated continually. This is the way to approach a balanced corpus. One should not try to make a comprehensive and watertight listing […] rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language […] In our ten years' experience of analysing corpus material for lexicographic purposes, we have found any corpus – however unbalanced – to be a source of information and indeed inspiration. *Knowing that your corpus is unbalanced is what counts* (Atkins et al., 1992: 10, italics mine).

Atkins et al.'s approach is attractive since it recognizes language as a growing and living entity which must be equally matched with a vibrant and growing corpus. Their position is shared by Čermák, who argues that,

> Thus it is hard to see why most (almost all) corpora are seen as strictly time-limited projects only which, when finished and having served their purpose, are far from being maintained, modernized, and substantially enlarged.... Since any language needs a consistent, perpetual, and next-to-exhaustive coverage of its data, it should have a corpus of corresponding qualities… This is particularly important in the case of minor languages which, unlike English and other languages, cannot afford the luxury of having a variety and multitude of corpora, at least not at the moment (Čermák, 1997: 182).

However both Atkins et al. and Čermák do not claim to have solved the matter of balance, rather they argue for a constant updating of the corpus over time – a position similar to that of Sinclair (1989: 29) who points out that "…a corpus should be as large as possible and should keep on growing". Even if a corpus is updated continuously, the challenge will remain in that some corpus linguists would want to work with a finite and constant entity such as the BNC rather than an entity whose contents are in perpetual flux.

It should be fairly clear that what constitutes balanced and representative corpora still remains controversial. The matter of how much sampling of different genres to include in a corpus is still largely unresolved. "The crux of the matter is finding a criterion for selecting the proportions between the reception and production" of text (Čermák, 1997: 192). What appears to be agreed upon though is that a corpus must finally capture the language varieties from a specified population from which a sample is taken, which reflects how that particular language community uses language. This is significant since (Summers, 1993: 186, 190) argues that the results of corpora analysis may be generalised to the general language community from which the samples were abstracted and Kennedy (1998: 94) shows the results of corpus analysis may have pedagogical function since "high frequency of occurrence as determined by the analysis of texts should be a major determinant of lexical content of language instruction".

Issues surrounding the exploration of linguistic variability have engaged many other researchers (Kittredge, 1982; Zwicky and Zwicky, 1982). Since corpora that substantially cover the full range of registers have been shown to be invaluable to both lexicographic research and studies in language variation, we are compelled that the corpus models for the Setswana language and other languages ought to represent a range of register diversity in both spoken and written situations.

4.3 Corpus annotation

Having collected texts into a corpus, such a corpus can contain simple raw text or it can be enriched with linguistic information before information extraction. The raw text can also be annotated or marked up. The mark-up language is concerned with the encoding of a corpus. The encoding, referred to as annotation or tagging, added to the texts that comprise a corpus, is a metalanguage that is generally done in some form of mark-up language (Horvath, 1999: Section 2.3.1). Two commonly used mark-up languages in corpora are XML and SGML. The Extensible Mark-up Language (XML) is the universal format for presenting structured documents and data on the World Wide Web (WWW). The functionality of the Web is improved through XML's design because it provides more flexible and adaptable information identification. "It is called extensible because it is not a fixed format like HTML (hyper-text mark-up language), which is a single, pre-defined mark-up language" (Pravec, 2002: 101). As a metalanguage, XML allows the design of customized mark-up languages for a limitless number of different types of documents. This is made possible because it is written in Standard Generalized Mark-up Language (SGML), the international standard metalanguage for defining descriptions of the structure for different types of electronic documents.

Grammatical tagging is one common practice of adding interpretative linguistic information to a corpus at various levels (Monachini and Picchi, 1992). It classifies each word-form in a text, labelling it with a part of speech tag (POS-tag) and morphological features. The process can be performed automatically. The part of speech mark-up is particularly crucial. De Rose (1991: 9) has shown that 11% of word types and 48% of word tokens occur with more than one category label (Kennedy, 1998: 209). For instance, the mark-up of the sentence: "There is nothing masculine about these new

trouser suits in summer's soft pastels." from the BNC (Burnard, 1995: 35) follows below:

<s n=00041>
<w EXO>There <w VBZ>is <w PNI>nothing <w AJO>masculine
<w PRP>about <w DTO>these <w AJO>new <w NN1>trouser
<w NN2-VVZ>suits <w PRP>in <w NN1>summer<w POS>'s
<w AJO>soft <w NN2>pastels<c PUN>.

The POS-tags in the above sentence are to be understood as follows:

AJO : Adjective
DTO : general determiner
EXO : existential there
NN1 : singular common noun
NN2 : plural common noun
PNI : indefinite pronoun
PRP : preposition, other than *of*
POS : the possessive or genitive marker *'s* or '.
VVZ : the –s form of lexical verbs, e.g. *forgets, sends, lives, returns*
PUN : any mark of separation (.!,:;-?..)
<s> : segment
<w> : word
<c> : a punctuation mark

The part of speech annotation can also be parsed or marked for syntactic information to show the phrase, clause or sentence divisions. The Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1994) is a sophisticated attempt at establishing guidelines of how to encode machine-readable text through a complex application of SGML. The SGML was used in the mark-up of the BNC which uses the Corpus Development Interchange Format (CDIF). This international standard provides, amongst other things, a method of specifying an application-independent document grammar, in terms of the elements which may appear in a document, their attributes, and the ways in which they may legally be combined (Burnard, 1995: 25). The detail of the mark-up is only relevant to the function to which the corpus would be put to as Kennedy (1998: 84)

shows: "The level of detail of mark-up has to be related to the potential use of the corpus." Programs such as CLAWS (Constituent Likelihood Automatic Word-tagging System) (Garside and Smith, 1997) have also been used in tagging various corpora like the BNC (see BNC website[11]).

.

Tagged corpora are useful in corpus linguistic research in that they can help in the development of disambiguation rules and facilitate automatic and semi-automatic syntactic analysis. Tagged corpora have also been found to be highly useful in the generation of word sketches. "Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004: 105).

Kilgarriff and Rundell (2001: 807) show that as corpora grow, so does the number of corpus lines for a word. This leads to what they call "the problem of information overload" for a lexicographer when he or she has to deal with a great number of concordance lines. The solution lies in statistical summaries. Kilgarriff and Rundell (2001) have generated word summaries through "Word Sketch" software which uses parsed corpus data to identify salient collocates – in separate lists – for the whole range of grammatical relations in which a given word participates (see also Kilgarriff and Tugwell, 2000). They report that lexicographers found that the Word Sketches not only streamlined the process of searching for significant word combinations, but often provided a more revealing, and more efficient, way of uncovering the key features of a word's behaviour than the method of scanning concordance lines. They offer detailed information that would be hard to extract from a corpus which is not annotated. We illustrate this with the word sketch for *pray* from Kilgarriff et al., (2004: 120).

Figure 2: Word sketch for pray (v)

pray **(v)** BNC freq= 2455

miracle	8	13.9	emperor	2	5.2	read	9	9.5	inwardly	3	5.5	hook	2	3.3	she	130	5.8
+for	**680**	**34.7**	**+to**	**142**	**4.1**	**and/or**	**679**	**7.7**	**+modifier**	**338**	**6.5**	**+object**	**183**	**3.2**	**+subject**	**1361**	**6.5**
forgiveness	12	19.8	spirit	32	4.30	singe	20	8.48	silently	15	4.43	girl	53	3.15	follower	306	5.03
soul	14	19.2	image	22	4.07	hopscotch	43	5.05	together	26	3.8	God	21	2.6	petitioner	3	4.8
me	247	17.3	lord	26	3.94	fast	6	3.92	fervently	4	3.6	pardon	6	2.6	knee	5	4.9
deliverance	61	16.6	saint	6	3.80	pray	56	3.52	loudly	6	3.5	day	2	2.4	congregation	4	4.8
peace	25	10.2	jesus	2	5.4	kneel	5	3.4	earnestly	10	3.3	silence	3	2.4	fellowship	263	4.0

[11] http://info.ox.ac.uk/bnc

church	12	11.7	believe	2	2.9	ever	9	3.0	Singh	2	3.7
guidance	8	11.6	learn	2	2.8	secretly	2	2.7	Family	6	3.6
us	16	11.6	tell	2	2.3	quietly	3	2.4			
chance	5	10.3				still	11	2.3			

The Word Sketch therefore helps reveal that people usually pray for *rain, soul, God, peace, peace miracles, forgiveness* amongst other things. It also reveals that the verb *pray* is usually modified by *silently, together, fervently, aloud* and *earnestly*. Such wealth of information would have been difficult to uncover without the help of Word Sketches.

4.4 Sample size

Every corpus is a language sample (Leitner, 1992). As discussed earlier (Chapter 3) a corpus can comprise sampled text from books, newspaper, speech and other text. Other corpora comprise complete works of writers, or complete texts such as the Bible, but they also in a sense constitute samples of language use by such writers or of particular genres. Such corpora will be discussed briefly later. What must be established foremost is that text sampling is central and basic to corpus construction. This position finds support in Biber, who points out that,

> Some of the first considerations in constructing a corpus concern the overall design: for example the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples within texts and the length of text samples. Each of these involves a sampling decision, either conscious or not (Biber, 1994: 377).

The matter of sample size is closely related to the previously discussed subject of representativeness since the number and size of texts in a corpus determine whether a corpus can be judged as representativeness of a language or not.

The purpose of sampling adequately is so that reliable generalizations may be made concerning a population as a whole. However, as we have seen, a linguistic population is normally so large (in terms of the number of speech acts produced) and so indefinable (in terms of the possible range of text types) that a random sample, stratified according to all major language text types, is probably not feasible (Kennedy, 1998: 74).

In corpus compilation one issue that still needs to be explored is how much of each text type sample should be included in a corpus. For those compiling opportunistic corpora, any amount of text found may be added to the corpus. For those attempting balanced corpora the need to define the population to sample becomes urgent and a decision of how much text from each text type must be made. However the language to be sampled, such as Setswana, as Clear (1992: 21) has argued, is poorly defined. Unlike in other studies where the population is clearly defined, say university students or people over the age of fifty, something like the Setswana language is not perfectly defined. It is broad with a variety of dialects; it is not clear whether we refer to produced (books, speech, etc.) or received language (language that we hear or read). It is also not clear what unit of language is best to be sampled and analysed, that is, whether we are interested in sampling words, sentences or whole texts such as books or conversations. The challenge that arises in sampling is that there is a real possibility that one may under-represent some variety of language in a corpus as Clear has shown:

> Given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample (Clear, 1992: 21).

Although defining a population to be sampled is difficult, it however has to be done if generalisations drawn from a corpus are to be made about a broad language community.

Different corpus compilers sample language differently. The Brown Corpus and the Lancaster Oslo Bergen (LOB) corpus each has 500 samples of 2,000 words each. Sinclair (1991: 19) argues that the even sample sizes are advantageous as far as making comparisons is concerned. In the BNC case a target sample of 40,000 words was chosen for books and anything less than 40,000 was reduced by 10% for copyright reasons (Burnard, 1995: 10).

Sinclair (1991: 19) points out that an alternative to smaller text samples is “to gather whole documents” and adopt a policy of continuous corpus growth since “from a large corpus can be drawn any number of smaller, more specialized ones, according to requirements from time to time.” The weakness of collecting whole documents as a collection strategy is that the coverage will not be as good as a collection of small

samples and one text characteristics may dominate others. On the size of a corpus sample, Biber (1995: 132) concludes that "1,000-word samples reliably represent many of the surface linguistic characteristics of a text, even when considerable internal variation exists."

Kennedy (1998: 20/21) argues that complete works corpus is "not representative of an entity. It is that entity."

De Haan (1992: 1) points out that one thing that has not been explored is how the size of corpus samples affects the research results. From a variety of experiments he conducts, he shows that the suitability of a sample depends on the specific study that is undertaken, and as if answering Biber's (1995: 131) question "What is the optimal text sample length?" he argues that there is no such thing as the best, or optimum, sample size.

Leech (1991: 10) argues that a preoccupation with size "…is naïve – for four reasons."

1. A collection of machine-readable text does not make a corpus. A corpus has to be designed for a specific representative function.
2. The vast growth of resources of machine-readable text has taken place exclusively in the medium of written language – speech devices have not developed the automatic input of spoken language to the level of the present OCR (optical character recognition).
3. While technology advances quickly, human institutions evolve slowly. Problems relating to copyright forbid the copying of text without the license of the copyright holder. It is therefore difficult to find corpus that is available unconditionally for all users.
4. While hardware technology advances, software technology lags behind. Having enormous amounts of text but lacking the software to explore them is unfruitful.

Leech shows that brute size in corpus compilation is not everything. The corpus must be representative; representing written as well as spoken language. He observes that developments in software technology will go far in aiding information retrieval from

corpora.

The brief discussion of sample size is aimed at showing that while sampling lies at the heart of corpus compilation, different corpus linguists adopt different sampling approaches. The Brown Corpus and the LOB corpus each has 500 samples of 2000 words each. The BNC comprises samples of 40,000 words for books and anything less than 40,000 has been reduced by 10% for copyright reasons (Burnard, 1995: 10). For those compiling opportunistic corpora, any amount of text found may be added to the corpus. It appears that the purpose to which a corpus would be used for need defining prior to any sampling. If a corpus is to be used to compare equal text samples then sampling chunks with equal number of words may be a desirable option. However in NLP, an opportunistic corpus may be ideal; while for lexicography, a corpus with broad coverage is desirable (see Manning and Schütze, 1999).

4.4.1 Spoken versus written corpus text

Speech in a language community is the primary channel of human communication and exists in abundance compared to written text (Cho and O'Grady, 1996). While this is common knowledge in linguistics, language researchers do not know quantitatively how much of speech exists, nor do they have the resources and methodologies to account for how many words are spoken daily by interlocutors.

General language corpora in order to better represent a language it must include both spoken and written text, different text genres and various dialectal varieties. If a corpus is compiled proportionally then spoken language would be greater than written language in a corpus. However this does not hold true in many corpora compilations since some are not sampled proportionally but in a stratified manner (see Section 4.2.1). Sinclair (2004) points out that "estimates of the optimal proportion of spoken language range from 50% — the neutral option — to 90%, following a guess that most people experience many times as much speech as writing." Such a greater occurrence of spoken text over the written one would approximate the ratios of written and spoken text in the real world and would be likely to produce corpora that closely represent language as used in speech communities. However in none of the large corpora like the BNC and the

Bank of English does the percentage of the spoken text exceed that of written text. The BNC, a 100 million words corpus of modern spoken and written English, has 90% written text and 10% spoken language. The ratios between the spoken and written corpus do not approximate the real world ratios of linguistic differences between spoken and written language. Sinclair (2004) argues that "most general corpora of today are badly balanced because they do not have nearly enough spoken language in them." This is true of the BNC although the BNC is one of the corpora with the largest spoken text (about 10 million words). Such an imbalance raises questions relating to the composition and balance of the corpus and also points to the fact observed by Sinclair (2004) that a corpus is an imperfect entity. He argues against any exactness in corpus compilation thus:

> It is important to avoid perfectionism in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like (Sinclair, 2004).

Leech et al. also recognise the inadequate representation of speech in the BNC thus:

> Although spoken language, as the primary channel of communication, should by rights be given more prominence than this, in practice this has not been possible, since it is a skilled and very time-consuming task to transcribe speech into the computer readable orthographic text that can be processed to extract linguistic information. In view of this problem, these proportions were chosen as realistic targets which, given the size of the BNC, are also sufficiently large to be broadly representative (Leech et al., 2001: 1).

According to Leech et al. the percentage of speech text in the BNC, was reached by determining what was possible to the compilers and not as a consequence of proportions of speech to written text in the English language. BNC designers could have arrived at the 90% and 10% ratios by studying the language situation of a speech community and projecting the estimated ratios of spoken and written language into the corpus structure. But according to Leech et al. these ratios were purely 'chosen as realistic targets' of limitations in the spoken language transcription and because of the expensive nature of manual transcription.

It is not clear if a situation in which a corpus has more spoken language is desirable for linguistic analysis. Biber (1994) has argued that to have greater spoken language percentages in a corpus is not linguistically interesting since the corpus ends up being homogeneous. What corpus compilers should aim for, he argues, are stratified corpora that capture the linguistic variability of the language community and not proportionally-compiled corpora. This position has however been rejected by Varadi (2001) who prefers proportional sampling and accuses Biber of attempting to redefine representativeness by divesting

> …such a key term of its well-established meaning, which has a clear interpretation to statisticians and the general public alike… There is such a strong and unanimous expectation from the public and scholars alike for corpora to be representative that it is an assumption that is virtually taken for granted. However, to meet this demand by the semantic exercise of redefining the content of the term is a move that hardly does credit to the field (Varadi, 2001: 592).

There are added challenges to spoken corpus compilation. It is not only the matter of what it means to be representative as seen in the different position taken by Biber and Varadi above. Atkins et al. express frustrations with building a corpus of spoken text when they say:

> The difficulty and high cost of recording and transcribing natural speech events lead the corpus linguist to adopt a more open strategy in collecting spoken language (Atkins et al., 1991: 3).

Atkins et al. suggest that technological inadequacies in speech transcription force corpus linguists to settle for corpora that are not desirable, but tolerable. Such positions inevitably raise the theoretical questions of whether corpus representativeness could be sustained in conditions in which the desired and representative corpora do not exist (see also Rundell, 1996).

Sharoff proposes one way of solving the lack of spoken material in a corpus in this way;

> The proposed solution is to increase the amount of ephemera (including leaflets,

> junk mail and typed material), correspondence (business and private) and spoken language samples whenever possible, because they reflect everyday language produced and reproduced regularly in discourse (Sharoff, 2004: 6).

Sharoff's attempts at solving the impasse illustrate the gravity of challenges of compiling spoken language. However the extent to which material such as business and private correspondence, leaflets and junk mail can substitute for spoken language is still to be investigated.

4.4.2 Newspaper text versus the purchase of a pair of shoes

Other researchers look at the matter of spoken language representation in a corpus differently. While they acknowledge the common occurrence of speech in daily discourse, they argue for more written language in a corpus since they consider speech private and restricted to a few interlocutors, while written text such as novels and newspapers have broad readership and deserve prominence in a corpus. One researcher who holds this view is Kennedy who argues:

> No one knows what proportion of the words produced in a language on any given day are spoken or written. Individually speech makes up a greater proportion than does writing of the language most of us receive or produce on a typical day. However, a written text (say in a newspaper article) may be read by 10 million people, whereas a spoken dialogue involving the purchase of a pair of shoes may never be heard by any person other than the two original interlocutors (Kennedy, 1998: 63).

Kennedy introduces an interesting dimension to corpus compilation that raises great controversy. It is true that a newspaper is likely to be read by many people and that its circulation may be verified from reliable sources. However the challenge still remains since newspaper buyers do not read the same sections of a newspaper. Some people have no time for business section, classified, cartoons, letters to the editor and other newspaper sections. Although circulation numbers might be available to assist corpus builders sample newspaper text, they only give numbers of purchased newspaper but do

not quantify patterns of newspapers readership.

A similar point may be made that although many of the corpora depend on published texts, there is indeed no guarantee that such texts are widely read (or read at all). This is particularly so in the Setswana language situation where the majority of Batswana do not read Setswana text, save in Setswana classes at both primary and secondary school. Kennedy (1998: 52) suggests that to fix this problem "best seller lists, library lending, statistics and periodical circulation figures can only partially reflect receptive use and influence." Kennedy's use of *partially* is an indication of the immensity of problems surrounding attempts to construct corpora on the basis of common and influential text. If "receptive use and influence" are taken as determinants of text inclusion in a corpus we must contend with varying *degrees* of such use and influence. School textbooks and creative texts read by thousands of students would be in use more than a library text that is rarely read. It is not clear how such a distinction will be reflected in corpus compilation. Textbooks would have been read more widely and therefore their text should somehow reflect the fact that they have been seen more than other texts. The argument may be pushed further. This would mean that a sign that reads: "Welcome to Gaborone" would make "welcome" "to" and "Gaborone" more common since such language would have been received by many people. However it is not clear how such information could be represented in a corpus. What about words like "Stop" used in traffic signs and seen by people repeatedly daily? Arguing for more of written language in a corpus since written language has been read widely or seen repeatedly compared to spoken language which is private, makes the discussion complex and in no way resolves challenges of the representation of speech in a corpus.

It would appear that Kennedy's argument against spoken text on the basis that it is private while written text is in the public domain, is not convincing but rather raises new problems and challenges as outlined above. Spoken text is as important as written text in corpus compilation and novel attempts need to be made to achieve its better representation in a corpus.

4.4.3 The value of spoken language

In this section we illustrate the value of spoken language to corpus research. We illustrate what might be missed if a corpus does not include spoken language text. Borrowings and colloquialisms are common in speech but they are dispreferred by editors and publishers, especially in communities where there is language contact such as the African context. Spoken Setswana is characterised by high levels of borrowing from English and Afrikaans. The documentation of foreign acquisitions in Setswana is not recent. Cole (1955) noted words like *beke* from "week", *baki* "*baadjie*" (jacket), *gouta* from "*goud*" (gold), *heke* from "hek" (gate), *hempe* from "*hemp*" (shirt), *kofi* from "*koffie*" (coffee), *pena/e* from "pen", *peipe* from "*pyp*" (pipe), *sukiri*, from "*suiker*" (sugar) and *baesekele* from "bicycle", *buka* from "*book*", *ofisi* from "office", *šeleng* from "shilling". There are other more recent borrowings like *gate* which reveal a certain layering in the nature of borrowed words. For instance, many Setswana speakers do not recognise *jase* from "jas" (coat), *heke* (hek) "gate" and *baki* (baadjie) "jacket", as borrowings from Afrikaans, while *jakete* (jacket) is recognised as borrowed from English. There is a similar situation with *heke*, which is considered by some speakers of Setswana as a sign of 'good old Setswana' while *geiti*[12] is recognised as an obvious borrowing. Spoken Setswana is peppered with instances of borrowing, code-switching and colloquialisms as illustrated in the following sentences.

Spoken Language	**English Equivalent**
Go shapo!	Bye
O tsile ka thelebišene	He came with a television
Ke bra/sistere ya gagwe.	It is his brother/sister.
O apere jase.	He is wearing a coat.

In the above examples *thelebišene* is a borrowing from the English noun *television* and *jase* from the Afrikaans noun *jas* and *shapo* a colloquialism which means *fine* or *bye*.

Borrowing and code-switching can also be seen in dialogue including days of the week, months and numerals. For instance, many Setswana speakers would say *Monday* or

[12] *geiti* is borrowed from the English "gate". Since Setswana does not have the voiced, velar plosive as part of its sound system, which in this instance occupies the initial word position in *geiti*, there is no agreed orthographic representation of such a sound in Setswana.

Mantaga (from Afrikaans, *Maandag), Saturday* or *Sateretaga* (from Afrikaans *Saterdag*) *and Sunday* or *Sontaga* (from Afrikaans, *Sondag*).

Setswana speech is also characterised by high degrees of code-switching, speakers switching from Setswana to English. This is particularly common in the use of English numerals in many instances instead of Setswana terminology. Many Setswana speakers would have difficulty in saying 1,567 in Setswana (i.e. *sekete, makgolo a matlhano le masome a marataro le bosupa*). Numbers are generally said in English. It is common for Batswana to use *one, two, three, fifteen, two thousand,* or *one million,* in their speech instead of Setswana terms *bongwe, bobedi, boraro, lesome le botlhano, dikete tse pedi* or *sedikadike*, respectively. Take the example below of a dialogue about selling. The example is from the spoken component of the Setswana corpus that we have compiled. English translations are given in brackets and numbers in Setswana speech have been italicised.

Dialogue 1

MT:	Shess... A a! ka nne ke letse ke bua le ene. A bo o mo neela ka *one fifty*. (Wow! But I was speaking to her yesterday. And you gave him for one fifty.)
TP:	*One fifty*?
MT:	Ee (Yes)
TP:	O ne a re wa re *sixteen* Pula. (She said you said sixteen Pula)
MT:	...Ke ne ke re, ka re *sixteen fifty*. (I was saying, I am saying sixteen fifty)
TP:	Ee, ke be ke mo neela ka *sixteen fifty*. Go tlhaela *six* Pula... (Yes, I then gave her for sixteen fifty. It is six Pula short.)
MT:	O a tlhaela? (It is short?)
TP:	Ee, a ke re ke ne ke mo tšhentšhetse ka madi ame. Ke raya gore ke tlaa tla ke mo go neela. (Yes, I gave her change using my money. I mean that I will give it to you later.)
MT:	Ehee. *Ok* nna ka re ke ena a sa, a sa, a sa mo ntshang. (Oh I see. OK. I thought that it was her who had, who had, who had not given the money.)
TP:	Nnyaa ao! Nnyaa. (No! No!)

From the above dialogue English numerals: *one fifty*, *sixteen*, *sixteen fifty*, and *six* are

used in the middle of a dialogue in Setswana instead of Setswana terms *lekgolo le botlhano, lesome le borataro, lesome le borataro le metso e e masome a matlhano* and *borataro* respectively.

It is not only English numerals which Setswana speakers usually switch to in speech, reference to months is also usually in English, and many speakers would have difficulties in stating months in Setswana. We return to this discussion later in this chapter.

Below we give two dialogues one a radio call-in program and the other from an interview television programme. The first two dialogues are from the Radio Botswana call-in program *A re bueng* (Let us talk) which is conducted largely in Setswana. The subject for the day was how certain youths abuse their parents by making difficult requests and demands, and if their demands were not met the youth threatened to commit suicide. We sample only a small part of the whole program. English words in the middle of Setswana dialogue are italicised and translations are in brackets.

Dialogue 2

RBP:	*Ok*, ba bangwe, o ise o tsamaye Mogotsi, ba re thupa ke yone (Ok, others, before you go Mogotsi, say whipping is the answer).
Cal:	*That doesn't solve anything and* mo go dira *to the worst* fa o…, ka na nna ke tle ke re le mo loratong a re e beye, fa o ratana le motho o bo a go raya a re: "Ke a go tlogela" O bo o re ke go rekela *something*… (That doesn't solve anything, it makes matters worse and if you can…, I sometimes say that in love relationships let us put it aside, if you are in a relationship with someone and they say to you: "I am leaving you" And then you say I am buying you something…)
RBP:	Mh.

Whole English sentences such as "[t]hat doesn't solve anything", phrases such as "to the worst" and words such as "ok", "and" and "something" are examples of the extent of English usage found in urban and educated Setswana speech. Below we give another speech chunk from the same call-in program.

Dialogue 3

RBP:	Fa gongwe e tlaa re a tsamaile a boe. (Sometimes after he leaves, he comes back.)
Cal:	A boe, mo ga go kgetla thupa o re ke betsa ngwana gore ga a batle go nkutlwa, *you are making things worse* go feta fa di leng teng. (He may come back, getting a stick to beat a child because he does not listen to me, you are making things worse beyond what they are).
RBP:	Nnya mme… (No but…)
Cal:	Thupa gotlhelele ga e yo tota le ko sekolong. *I don't encourage,* gore ba re thupa e ka sokolola ngwana. *Sit down* le motho, buang le ene o tlaa ipaakanya. Fa go pala go raya gore go a pala. (Whipping completely is not there at school. I don't encourage, that they say that whipping can transform a child. Sit down with someone, speak to them, they will fix themselves. If it fails, it would have failed).

Similar to the previous speech chunk investigated above, English sentences creep into Setswana speech. For instance: *You are making things worse* and *Sit down*. There are also clauses such as *I don't encourage*. We need to keep in mind that radio call-in programmes are informal programs where callers freely express their views on a variety of issues. We will however see that even in formal programmes a similar pattern of switching to the English language persists.

We now look at a formal television programme broadcast in the Setswana language. While participants in this programme come prepared to address a specific subject, they do not know the questions in advance.

The following dialogue was transcribed from the Botswana television programme, *The Eye*, which is an interactive programme with two to three interviewees tackling a current matter of concern. The subject of the program was on the drying Gaborone dam which supplies the capital city with water and the role of the Botswana Water Utilities and Water Affairs in advising and training users in water conservation.

Dialogue 4

OS:	Mme se re tshwanetseng gore re se gakologelwe ke gore jaaka Mma SR a ne a bua kgantele ka gore metsi a mo matamong a kgadisiwa ke, ke *evaporation* go na le *elemente* e nngwe gape e e leng gore e teng ya gore, letamo jaaka o le itse le nna le ... mmu jaaka o ntse o tsena mo letamong o fokotsa *capacity*… (But what we should remember is that as Mrs. SR was saying earlier that water in the dam dried because of evaporation, there is another element at play, which is, the dam as you know has… as soil collects into the dam it decreases the capacity…)
MK:	*So* re lebile (So, we are looking at) *eight months as the best case scenario, worst case scenario*?
GS:	*Worst case scenario* mma tota re ka nna ra re (Worst case scenario, we can say) *between six and eight months*.

Dialogue 4 shows a formal educated dialogue characterised by words such as *evaporation*, *element* and *capacity* and phrases such as *worst case scenario* and *between six and eight months*. English is pervasive in spoken Setswana as Bagwasi (2003) has shown.

There are also cases of colloquialism in spoken language. An example of colloquial speech from the Setswana corpus follows (English words are bolded and colloquialisms italicised):

> Hey monna Bobi, o seka wa dira *daidee*. *Magents* bane ba tseela Tshege dilwana *daa*, a *vaela dladleng* a le maponapona, **fortunately** bane ba sa nne kgakala *plus* **it was at night.** Hey phikwe, re *chitse* ha posong *baba* gongwe ko statung (statue) rena le Comfort a nwa coca cola, **Saturday afternoon**, re planela maitseboa. Re bo re *shapa round* mo *mmolong*, re o *covera* **in 10 minutes**.O *vaa* ka **line** ya Elegant, *ga* otla o tswa ka ko Pep kakwa otla ka **line** ya Pioneer town e fedile, heish *Zana baba*.

> Hey man Bobi, don't do that. Guys stole Tshege's clothes at that place and he went home naked, fortunately they did not live far and it was at night. I remember Phikwe, we relaxing next to the post office or the statue together with Comfort drinking a Coca Cola on Saturday afternoon planning for the evening. Then we would go around the mall and cover it in 10mins. You would go from the Elegant side coming from the side of Pep stores, the side of Pioneer and you would have covered the entire mall. How I miss Phikwe! (*translation mine*).

In the above quoted text *baba* (man, sir), *shapa round* (leave and return quickly), *mmolo* (mall), *covera* (cover), *vaa* (go), *daidee* (that thing), *magents* (guys), *chitse* (chilling, relaxing), *vaela* (go towards), *daa* (there), *dladleng* (home) are all colloquial Setswana words which are not used in formal texts. It is in analyzing spoken language that the colloquialisms are encountered. The presence of colloquialisms in speech lends additional support to the inclusion of transcribed spoken language in a corpus.

What we have attempted to show so far with the different dialogues and an example of colloquialisms is that the entity called Setswana spoken language is not a uniform, clean and homogeneous phenomenon. Rather it is characterised by foreignisms and colloquialisms. Borrowing, colloquialisms and code-switching are therefore some of the issues which confront Setswana lexicographers who use a Setswana spoken corpus or a corpus comprising portions of spoken data. Such lexicographers would grapple with issues relating to spoken text amongst these being:

1. The transcription of the language. Apart from it being a time-consuming process, there are tough decisions to be made on what is borrowing and what is merely code-switching.
2. If the corpus is annotated, there will be decision on what to mark-up (coughs, sneezes, passing traffic, hesitations, etc).
3. At a practical lexicographic level some of the issues that arise from including transcribed spoken language in a corpus include decisions of the kind of borrowed words to be listed in the dictionary and the kind of stylistic information derived from borrowed words.
4. The spelling of certain words on which there is no agreement.

5. Speech which is not thought through, characterised by hesitations, back-tracking and incomplete sentences.

The challenges of the treatment of borrowings in dictionaries that face a Setswana lexicographer mainly because of spoken text in a corpus are not unique to the language. Another language that faces a similar challenge is Toqabaqita, an Austronesian language spoken in the Solomon Islands.

The inclusion of spoken language in a corpus has relevance to the treatment of code-switching and borrowed words abstracted from such a corpus in a dictionary. In the subsequent section we discuss how lexicographers have addressed the challenges of borrowing and code-switching in the Toqabaqita language and how their approach sheds light to the treatment of borrowings and code-switching to the Setswana language.

4.4.4 The treatment of borrowings in Toqabaqita

Because of language contact many languages borrow words form others. This raises questions of whether such borrowed words qualify as belonging to the borrowing language and therefore deserving to be in its dictionaries. Lichtenberk (2003) in his report on the dictionary of Toqabaqita points out that the central point in determining the wordlist of a dictionary is the consideration of intended users of a dictionary, what he calls “audience”, and expectations, that is, the kind of purpose the dictionary has to serve in the society. This view is shared by Zgusta who says decisions of what to include are determined by “fundamental decisions concerning the type of dictionary which is to be prepared” (Zgusta, 1971: 243). For instance if the dictionary intends to contribute to historical and comparative studies it may list archaic and obsolete words while the inclusion of loanwords may prove to be of interest to phonologists. But the larger part of Lichtenberk’s (2003) paper is devoted to the discussion of inclusion or exclusion of loanwords in the dictionary of Toqabaqita. We discuss it in detail since there are comparisons which may be drawn between Toqabaqita and Setswana. Lichtenberk is confronted with a language situation where he has to make a decision of whether to include Pijin words in the dictionary of Toqabaqita since some of them fit the phonological and phonotactic constraints of Toqabaqita while others do not. Like

Setswana, Toqabaqita does not permit consonantal cluster or syllable final consonants and has a simple syllable structure of CV and V. This is exemplified in words like *kisini*, "kitchen" and *wasia* "wash". The principle that guides Lichtenberk in deciding what to include is:

> Pijin words used in Toqabaqita are listed provided they fit the phonological and phonotactic patterns of Toqabaqita, either because they fit them already in Pijin or because they have been accommodated to them. Words which do not fit the patterns are not listed (Lichtenberk, 2003: 395).

This principle excludes certain words that are in common use which in Lichtenberk's view are instances of code-mixing (Lichtenberk, 2003: 396) and not borrowing. These words include *qambrela* "umbrella" from Pijin *ambrela* and *grup* or *grupu* "group" from Pijin *grup*. They are not listed in the dictionary since they do not satisfy the phonotactic constraints of Toqabaqita. Similar to the Setswana situation, code-mixing in Toqabaqita is common, especially in numerals, months and the names of some of the days of the week and Lichtenberk argues:

> Considering such words to be part of Toqabaqita lexicon would amount to claiming that the phonological inventory and the phonotactic patterns of the language have undergone some major changes (Lichtenberk, 2003: 396)

Therefore Lichtenberk decides to restrict the matter of code-mixing to the front matter where the common but non-accommodated words would be listed. There are also problems concerning pairs of words which though accommodated from Pijin, have variants which do not conform to the phonotactics of Toqabaqita. In this instance the variant that does not conform to the phonotactic constraints is not listed. This is exemplified by *bereta* and *bret* "bread" where *bereta* is accommodated and *bret* is not listed since it is less common and not accommodated. The situation gets increasingly interesting when the non-accommodated variant is more common than the accommodated one as in *gavman* (that violates the phonotactic constraints of Toqabaqita and is un-accommodated) and *gafumanu* (is accommodated but it is infrequent). In such a case Lichtenberk ignores the most frequent used word *gavman*, since it violates the phonotactic constraints of the language, and instead chooses to enter

the less common *gafumanu* on the principle that the non-accommodated variant though frequent, is an instance of code-mixing.

Lichtenberg develops other principles which govern what to list, and these are listed below:

1. "Words that belong in well-circumscribed and relatively small sets are not listed if some other members of the same set do not occur in an accommodated form and so are not listed" (Lichtenberk, 2003: 396). Such sets include numerals, days of the week and names of months.
2. A Pijin word that has been encountered only once is not listed even if it fits the phonological and phonotactic pattern of Toqabaqita.

The question of what has to be listed in the dictionary raises an issue of the boundaries of the lexicon of a language. And Lichtenberk divides the Toqabaqita into 3 categories: i) native Toqabaqita words ii) accommodated borrowings from Pijin, and iii) Pijin words used without being accommodated. Lichtenberk concludes that:

> Only the first two types are to be listed in the dictionary, which amounts to saying that only those words are part of Toqabaqita lexicon, while the non-accommodated words are not (Lichtenberk, 2003: 397).

And Lichtenberk gives proper criticism to his approach when he says:

> The principle, while explicit and applicable in a straight forward way, is nevertheless arbitrary. It gives priority to the phonological and phonotactic patterns of Toqabaqita over usage. Pijin words that are not accommodated are, by fiat, placed outside the circumference of the Toqabaqita lexicon, although by virtue of their usage they could be inside (Lichtenberk, 2003: 397).

Lichtenberk's criticism of his principles is accurate. His principles could lead to unacceptable results. Take for instance the principle that: "Words that belong in well-circumscribed and relatively small sets are not listed if some other members of the same set do not occur in an accommodated form and so are not listed" (Lichtenberk, 2003:

396) which include numerals, days of the week and names of months. While this principle might work well in reference to numerals and names of months in Setswana, the same cannot be said for days of the week. Let us consider the days of the week data in Setswana:

Table 16: Setswana days of the week

English	**Standard/written**	**Kgasa (1976)**	**Spoken/Common**
Sunday	*Tshipi*	*Lantlha (Tshipi)*	*Sontaga*
Monday	*Mosupologo*	*Labobedi*	*Mantaga*
Tuesday	*Labobedi*	*Laboraro*	*Labobedi*
Wednesday	*Laboraro*	*Labone*	*Laboraro*
Thursday	*Labone*	*Labotlhano*	*Labone*
Friday	*Labotlhano*	*Laborataro*	*Labotlhano*
Saturday	*Matlhatso*	*Labosupa (Sabata)*	*Sateretaga*

Table 16, shows days of the week in Kgasa (1976), in common spoken language and in standard written Setswana. Standard Setswana names are used in text books, novels, and government media and in creative writing in schools. In the table the column with standard Setswana is followed by a recommendation of the days of the week by Kgasa (1976) in the front matter of the Setswana dictionary. His list is a purist approach of avoiding borrowings from Afrikaans as he says:

> Malatsi a beke (tshipi) a ka bidiwa ka Setswana ka motlhofo go sena Sekgowa le fa e le Seburu (Kgasa 1976: front-matter).
>
> [Days of the week can be referred to easily without resorting to English or Afrikaans (*translation mine*)].

In the above quotation Kgasa is at pains in shrugging off borrowings but even the very Setswana sentence he uses to shun Afrikaans, has at least two borrowings from Afrikaans. These are *beke* 'week' and *Seburu* from 'Boer'.

Additionally, Kgasa rejected certain names of days of the week in standard Setswana such as *Matlhatso* which he considered to be religiously insulting to others. He objected that:

> Fa malatsi a bidiwa jaana ga gona nyenyafatso ya tumelo ya ba bangwe ka lefoko la Matlhatso jaaka go ntse gompieno (Kgasa, ibid)
>
> When the days of the week are referred to this way (in the way he suggested) there is no condescension of other people's faith with the term Matlhatso (Saturday) as it is today (*translation mine*).

Matlhatso is a noun derived from *tlhatswa* 'wash' and Kgasa may have perceived the name to be offensive to the Seventh Day Adventists (SDA) who consider Saturday as a day of rest and not for manual labour such as washing. Kgasa also objected to the use of the name *Mosupologo*:

> Lefoko la Mosupologo ga le utlwale ka gobo (sic, *go bo*) tota beke e a bo e sa robala mo e reng letsatsi le le salang Lantlha morago le bo le bidiwa Mosupologo jaaka ekete beke e a supologo (sic, *supologa*) (Kgasa, 1976: front matter).
>
> The word Mosupologo does not make sense because a week is not asleep, such that the day after Sunday should be called Mosupologo as if a week rises from dust (*translation mine*).

Kgasa understood that the noun Mosupologo is derived from the verb *supologa* 'rise from dust' and he found this inaccurate to refer to a day at the beginning of the week. But he was too late; the word had caught on and his recommendation never gained currency. His suggestion only jumbles the names of days of the week resulting with Monday called Tuesday (see Table 16). This failed attempt by Kgasa approximates Churchward's (1959) inventions of loan words in his dictionary (see Lichtenberk, 2003: 394).

What is surprising concerning Kgasa's recommendations is that Setswana authors before him did not share his views. For instance, Sandilands (1953: 153) days of the week are dissimilar to Kgasa's recommendations:

Table 17: Sandiland's rendering of days of the week

Setswana	English
Lamoréna[13], Tshipi	Sunday
Mantaga, Mosupologò	Monday
Lwabobedi	Tuesday
Lwaboraro	Wednesday
Lwabonè	Thursday
Lwabotlhano	Friday
Matlhatsò, Maapèò[14], Satertaga	Saturday

Although some of the terms used by Sandilands have since gone out of usage, his rendering of days of the week is closer to the way Setswana is currently spoken compared to Kgasa's recommendations.

But of immediate relevance to this section also is what we list as Spoken/Common names of the week. The list includes borrowings *Mantaga/Mmantaga, Sateretaga,* and *Sontaga* from Afrikaans *Maandag, Saterdag* and *Sondag* respectively. Contrary to Lichtenberk's recommendations, excluding these borrowings from a Setswana dictionary would make it highly deficient since they are common in spoken language and increasingly used in the media, parliament and other domains of Setswana language use as illustrated in the concordance lines below.

Figure 3: Mantaga concordance lines

```
1      e jaaka ekete ke tsatsi la Sontaga. Mantaga mongwe le mongwe thupa e n
2      eng thata ka metlae, e leng Luzboy, Mantaga le Laboraro mongwe le mong
3      o ga a site go sita loso. E ne e le Mantaga thapama fa Motsei a tswale
4      wa sebining ya ga Motsei. E ne e le Mantaga mme nako e ka nna ya bosup
5      ka pampiri (di-mask). 45 Lenaneo la Mantaga - Std 4 Bana ba ithuta ka
6      e dilo tsa gago tsa go ya tirong ka Mantaga. , : Mosadi o o jaaka wena
7       se tima. Fa rraagwe a ya tirong ka Mantaga, a gakgamatswa ke fa sejan
8      olo ya gore o tl ya kwa teropong ka Mantaga a ye go reka dipampiri go
9      eleng ba ne ba tla boela Tembisa ka Mantaga thapama. Bana ba ga Daphne
10     ile phitlhong pele ga e sutisiwa ka Mantaga, Mogokgo wa sekolo se sego
11     sigo ka Satertaga le ka Sontaga. Ka Mantaga o ne a tshwarwa ke dipapal
12     go go itsise gore o tla simolola ka Mantaga. : Ke tla kgona go ya tiro
13     RONE: Erile Palamente e simolola ka Mantaga, T ona ya T emo-thuo, Dani
14     ng lengwe le lengwe, go simolola ka Mantaga go ya kwa go Labotlhano. B
15     teretaga le erne jaana: simolala ka Mantaga — Sateretaga 6 a.m. ke nak
16     e 4 se ka a itse go tla sekolong ka Mantaga. Re a bona jaaka mosetsana
17     ka rakana le Mosela kwa sekoleng ka Mantaga. Re ne re na le boikutlo b
18     senwa. Go tloga fa re ya gae mme ka Mantaga o tla mpolelela maina a di
```

[13] The use of this word to refer to Sunday has almost disappeared from Setswana use and may only be found amongst very few old speakers of Setswana, in very rare occasions.

[14] The use of this word to refer to Saturday is no longer in current Setswana usage.

```
19    ne e le mafelo a beke a maleele, ka Mantaga e ne e le letsatsi la boik
20    a go tlhatswa dikhai tsa Makgowa ka Mantaga le ka Labobedi mme a be a
```

Figure 4: *Sontaga* concordance lines

```
1     ne yo o neng a tshaba go lema yole. Sontaga mongwe le mongwe bana ba d
2     botsa Mmadisenke mo tshokologong ya Sontaga ba robile sogo tno phaposi
3     esele! Ga ke tshoswe ke modumedi wa Sontaga fela. Mo bekeng re a tshwa
4     " Ga bua Kepaletswe monyebo e le wa Sontaga, "Dumela Kepaletswe. Ke en
5     tshameko ya bosheng, bogolo jang wa Sontaga mme re ka nametsega re le
6      tshipi. Ke t/aa go bona ka Tshipi (Sontaga). Ba tlaa goroga tshipi(be
7      sa ga Mosela mo nokeng e Tshetlha. Sontaga e e latelang go ne ga bewa
8     na f~a a laela. "Ke tla go tshakela Sontaga se se tlang ..." A bua a
9     se golo mo re ja dilo,le dipina tsa sontaga , gape ke batla go bona ba
10     sonola sonolega sonolegile sonotse Sontaga sonya sopaladitse sopalala
11    mo sonobolomo sonobolomo sonobolomo Sontaga Sontaga sontile soutile so
12     e telele. Kag~so o ile sekolong sa Sontaga le mme. Mme o farile Kagis
13    a Sontaga dipina. Bana ba sekolo sa Sontaga ba ntse mo ditilong. Ba op
14    tseboa 8.30 p jn. nako ya go robala Sontaga 7.00 a.m. nakoya go tsoga
15    majana a a lesome le bosupa. E rile Sontaga kefa lonyalo Iwa rona lo b
16     ffe yo Bham, o bula Sateretaga, le Sontaga tota o kgona go thusa bath
17    mo mafelong a beke, ka Matlhatso le Sontaga, fa a sa ya go bogela mots
18    eme leganbng. Go ne go le tsatsi la Sontaga, Ka nako tsa lesorne mo tl
19    lhela ka Sateretaga. Ka letsatsi la Sontaga ba lelwapa ba ne ba ya
20    a tlhomamiso ba tla tlhomamisiwa ka Sontaga. 6. Lokolola polelonolo e
```

Such names of the week could be marked in a dictionary as *common in spoken language*, or as *colloquial*. But it would be unsatisfactory not to list them in a Setswana dictionary just because a small set (of Afrikaans names of the week) from which they are derived, is not borrowed into the Setswana language in its entirety. Frequency here should be considered paramount.

The Setswana dictionaries have treated the different three borrowing in different ways. Brown (1925) does not enter *Mantaga*, *Sateretaga* and *Sontaga*. Kgasa (1976) enters *Mantaga* and not *Sateretaga* and *Sontaga*. Snyman et al (1990) include *Sontaga* and *Mmantaga* in the dictionary but leave out *Sateretaga*. Matumo (1993) does not enter *Mantaga*, *Sontaga* and *Sateretaga*. Kgasa and Tsonope (1998) enter *Sontaga* and not *Mantaga* and *Sateretaga*.

Word frequency lists are helpful in decisions of what to enter in a dictionary. Listing frequent borrowings such as *Sontaga*, *Mondaga* and *Sateretaga* and marking them as either colloquial, belonging to spoken language or as foreignisms would be a preferred approach.

Obviously the kind of dictionary being built would influence such decisions; whether it is monolingual or bilingual, intended for learners or for general use, or whether it is a dictionary of slang or not, primarily for encoding or decoding (e.g. academic use, which is a different case) and the number of pages a lexicographer has to work with.

Additionally, cases where certain terms, though known in the native language are rarely used in speech, but are replaced by borrowings and code-switchings, cannot be ignored (cf. Otlogetswe, 2006). This is particularly true for numerals where one finds sentences like, *O rekisitse dinamune di le* ten. "He sold ten oranges". *Mmiting o ka* ten *kamoso.* "The meeting is at ten tomorrow". In these examples, the speaker has chosen the English word *ten*, instead of the Setswana term *lesome/some*. The transcription of the term *ten* as either *ten* or *thênê*, as in the above examples, is based on the theoretical question of whether such a term has gained currency as an instance of borrowing or of code-switching. Are lexicographers to assume that such language usages do not exist in the language and that they do not have any relevance to dictionary compilation? Any answer to this question would lead to disagreements between lexicographers.

A similar pattern may be observed in days of the week with *Sateretaga* (Saturday), *Sontaga* (Sunday), *Mantaga* (Monday), and *wikente* (weekend) being more colloquial and common in spoken language than in the written form while *Matlhatso* (Saturday), *Tshipi* (Sunday), *Mosupologo* (Monday) and *mafelo-a-beke* (weekend), are common in written text, formal address and amongst the elderly. The stylistic information is significant particularly in dictionaries that attempt to achieve a fuller understanding of a word's meaning and usage. When both formal and informal terms are included in a dictionary, they may provide valuable stylistic information and may also be significant to future research as to when a word entered the language or when it changed its meanings.

This hopefully shows the importance of including greater occurrences of spoken text in a corpus since spoken language is used more in human communication and possesses unique characteristics not common in written language.

Next, the design of the two English corpora is considered.

4.5 Brown Corpus and BNC review

In Chapter 5 we discuss the Setswana corpus design and compilation. Before that we review two corpora which have been influential in English corpora analysis: The Brown Corpus and the BNC.

4.5.1 The Brown Corpus

Corpus linguists usually make reference to the *Brown University Standard Corpus of Present-Day American English*, commonly known as the Brown Corpus, (Francis and Kucera, 1964) as having pioneered research in corpus computational linguistics. The Brown Corpus was "significant not only because it was compiled for linguistic research, but also because it was compiled in the face of massive indifference if not outright hostility from those who espoused conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky" (Kennedy, 1998: 23).

The Brown Corpus was compiled by Nelson Francis and Henry Kucera in 1961. The corpus has over a million tokens of written text published in the USA in 1961. The Brown Corpus comprises 500 samples of about 2,000 tokens of continuous written English which approximate 1,014,300 tokens. Table 18 gives the text categories of the Brown Corpus and the proportions of different portions of the corpus.

Table 18: Structure of the Brown Corpus

Text type	**Proportion**
i. Informative Prose	**75%**
A. Press: Reportage e.g. Political, Sports, etc	8.8%
B. Press: Editorial e.g. personal, letters to ed., etc.	5.4%
C. Press: reviews e.g. books, music etc.	3.4%
D. Religion e.g. tracts, books, etc.	3.4%
E. Skills & Hobbies e.g. periodicals, books, etc	7.2%
F. Popular lore	9.6%

e.g. books, periodicals, etc	
G. Belles letters, biography, memoirs etc	15%
H. Miscellaneous e.g. government documents, industry reports, college catalogue, etc.	6%
I. Learned e.g. medicine, mathematics, law, etc.	16%
ii. Imaginative Prose	**25%**
J. General Fiction Novels and short stories	5.8%
K. Mystery and Detective Fiction Novels and short stories	4.8%
L. Science Fiction Novels and short stories	1.2%
M. Adventure and Western Fiction Novels and short stories	5.8%
N. Romance and Love Story Novels and short stories	5.8%
O. Humour Novels and essays, etc	1.6%

According to Kucera and Francis (1967: xvii) the samples were selected by "a method that makes it reasonably representative of current American English".

Ide and Macleod (2001: 274) argue that while the Brown Corpus has been extensively used for natural language processing work, its million words are not sufficient for today's large scale applications. For example, for tasks such as word sense disambiguation, many word senses are not represented, or they are represented so sparsely that meaningful statistics cannot be compiled. Similarly, many syntactic structures occur too infrequently to be significant. The Brown Corpus is also far too small to be used for computing the bigram and trigram probabilities that are necessary for training language models used in a variety of applications such as speech recognition. Fillmore et al. (1998: 966) have also found the Brown corpus to be "too small to provide adequately large samples for the purposes of lexicon construction."

Furthermore, the Brown Corpus, while balanced for different written genres, contains no spoken English data. Ide and Macleod (2001) lament the fact that while the 100 million words of the BNC provide a large-scale resource and include spoken language data; it is not representative of American English. As a result, there is no adequate large corpus of American English available to North American researchers for use in natural language and speech recognition work. Ide and Macleod (2001), because of this lack have argued that there is a need for a corpus of American English that is similar to the

British National Corpus. The project to compile the American National Corpus comparable to the BNC is detailed in Ide et al. (2002). They have shown that there are significant lexical and syntactic differences between British and American English. They point to the well-known variations such as: "at the weekend" (Br.) vs. "on the weekend" (U.S.), "fight (or protest) against <something>" (Br.) vs. "fight (or protest) <something>" (U.S.), "in hospital" (Br.) vs. "in the hospital (U.S.), "Smith, aged 36,…" (Br.) vs. "Smith, age 36…" (U.S.), "Monday to Wednesday inclusive" (Br.) vs. "Monday through Wednesday" (U.S.), "one hundred and one" (Br.) vs. "one hundred one" (U.S.), etc. Also, in British English, collective nouns like committee", "party", and "police" have either singular or plural agreement of verb, pronouns, and possessives, which is not true of American English.

Rayson and Garside report that the Brown corpus has been used in one of the largest comparative studies of the one million words of the American English (the Brown corpus) with one million words of British English (LOB corpus) by Hofland and Johansson. (1982). They also report on Yule's (1944) coefficient measurement which showed the relative frequency in the two corpora. Kilgarriff (1997a) used the Brown corpus to measure corpus homogeneity. The Brown corpus has also been studied for the abstraction of collocations. It has been found that the Brown Corpus has only two instances of "cups of coffee", five of "for good" and seven of "as always" (Kjellmer, 1994a).

The Brown corpus has therefore been a useful resource for linguistic research. However as has been seen, it was just too small for studies which needed large corpora. One corpus which was compiled to respond to this need is the British National Corpus.

4.5.2 The BNC review

The BNC is a 100 million-word corpus of written and spoken language from a variety of sources, designed to represent a wide cross-spectrum of current British English. The corpus "contains just over 4,000 texts" (Aston, 2001: 73). It was compiled by by a consortium of dictionary publishers and academic researchers between 1990 and 1994. These included the Oxford University Press, Longman Group Ltd, Chambers Harrap,

Unit of Computer research on the English Language (Lancaster University), Oxford University Computing Services, and the British Library Research and Development Department. Ninety percent of the BNC are written texts while 10% of the BNC is transcribed spoken text.

The BNC compilation was funded over three years with a budget of over GBP 1.5 million. The project was funded by the commercial partners, the Science and Engineering Council (now EPSRC) and the DTI under the Joint Framework for Information Technology (JFIT) programme. Additional support was provided by the British Library and the British Academy (see the BNC website: http://www.natcorp.ox.ac.uk/).

4.5.2.1 The BNC design criteria

Since the BNC was compiled so that generalizations could be made on the British English it was crucial that varieties that existed in the British English be represented in the corpus. The BNC was therefore built by sampling materials from across the language with respect to explicit design criteria rather than basing the collection of texts on their availability. Burnard notes that,

> The objective was to define a stratified sample according to stated criteria, so that while no-one could reasonably claim that the corpus was statisticxally representative of the whole language in terms either of production or reception, at least the corpus would represent the degree of variability known to exist along certain specific dimensions, such as mode of production (speech or writing); medium (book, newspaper, etc.); domain (imaginative, scientific, leisure, etc.); social context (formal, informal, business, etc) and so on (Burnard, 2002: 21).

The BNC design criteria specify a range of text characteristics and proportions for the material to be collected (see Atkins, 1992). Below we briefly look at both the written and spoken language design criteria of the BNC.

4.5.2.2 The BNC written component

Ninety percent (89,740,544 words) of the BNC is written texts that were classified into two principal parallel categorisations of:

a. *domain* (i.e., subject matter, divided into nine classes, viz., imaginative; arts; belief and thought; commerce; leisure; natural science; applied science; social science; world affairs: from 146 to 527 texts in each), and
b. *medium* (five classes, viz., book; periodical; miscellaneous published; published; to-be-spoken: from 35 to 1,414 texts in each). All the texts were selected on the basis of a publication period, marked as *time* in the corpus (Aston, 2001: 73).

The written part includes extracts from regional and national newspapers, specialist periodicals and journals for different ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.

The criterion of *domain* refers to the content-type of the text; *time* refers to the period of text production, while *medium* refers to the type of text publication, as in newspaper or book. Table 19 summarises the contents of the three criteria (see Aston and Burnard, 1998: 28-33).

Table 19: The BNC written components

Domain	**%**
Imaginative	21.91
Arts	8.08
Belief and thought	3.40
Commerce and finance	7.93
Leisure	11.13
Natural and pure science	4.18
Applied Science	8.21
Social Science	14.80
World Affairs	18.39
Unclassified	1.93
Time	**%**
1960-1974	2.26
1975-1993	89.23
Unclassified	8.49
Medium	**%**
Book	58.58
Periodical	31.08
Misc. published	4.38
Misc. unpublished	4.00
To-be-spoken	1.52
Unclassified	0.40

4.5.2.3 The BNC spoken component

The design of the spoken component of the BNC adopted a two-part approach: demographic and context-governed. The demographic approach employed demographic parameters to sample everyday speech of the British English speakers in the United Kingdom. The context-governed approach attempted to cover the full range of linguistic variation found in spoken language using a typology based on four contextual categories: educational (lectures, news broadcasts etc), business (sales demonstrations, union meetings etc), public/institutional (sermons, political speeches etc) and leisure (sports commentaries, radio phone-ins etc) (Crowdy, 1994). The demographic component, on the other hand, comprises recordings of 124 volunteers from four different social classes, male and female, different age groups and various geographical regions.

The spoken component constitutes 10% (10,365,464 words) of the BNC. For the spoken component, a first distinction was between "demographic" (conversations: 153 texts) versus "context-governed" (speech recorded in particular types of setting: 757 texts), and the "context-governed" component was further divided according to the nature of the setting (educational/informative; business; public/institutional; leisure: from 131 to 262 texts in each), paralleled by a monologue/dialogue distinction (40%/60%) (Aston, 2001: 73). Table 20 summarises the divisions in the corpus. It covers both the demographic and context-governed components and the context-governed component structure.

Table 20: The BNC spoken components

Context-governed	**%**
Leisure	23.71
Institutional	21.86
Business	21.47
Educational and Informative	20.56
Unclassified	12.38
Region	**%**
South	45.61
North	25.43
Midlands	23.33
Unclassified	05.61
Interaction type	**%**
Dialogue	74.87

Monologue	18.64
Unclassified	06.48

The value of compiling such a stratified corpus was to try and capture the varieties of modern British English from the 60s until the early 90s. It was designed to characterise contemporary British English "in its various social and generic uses" (Aston and Burnard, 1998: 28). Such linguistic variability was crucial for the corpus so that authoritative generalisations about the language could be made confidently. This need for compiling representative corpora from which generalisations could be made and on which hypothesis could be tested is expressed by Renouf thus:

> When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested. The first step towards achieving this is to define a whole of which the corpus is to be sampled (Renouf, 1987: 2).

The BNC has been useful for a wide variety of language research purposes including dictionary compilation of the *Longman Dictionary of Contemporary English* (3rd edition) (Summers, 1995), *Oxford Advanced Learner's Dictionary of Current English* (Hornby, 1996), *Longman Essential Activator* (1997) and *The New Oxford Dictionary of English* (Pearsall, 1998). The BNC "was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography" (Kilgarriff, 2001: 342).

Leech et al. (1997) explored the social differentiation in the use of English vocabulary in the BNC while Čermák and Kren (2005) compare its composition with that of Czech National Corpus. Rayson et al., (1997) undertake selective quantitative analyses of the demographically-sampled spoken English component of the British National Corpus. They compared the vocabulary of speakers according to gender, age and social group. The BNC has also inspired the compilation of other corpora such as the American National Corpus (Fillmore et at., 1998), the Russian Reference Corpus (Sharoff, 2004) and the Czech National Corpus (Čermák, 1997).

4.6 The exploration of both corpora

After a corpus has been compiled, "lexicographers need the skills and/or the software to navigate through sometimes huge numbers of corpus instances" (Kilgarriff, 2000: 109). However it has been found that there is a lack of tools for corpus-based lexicography, especially in relating corpus observations to dictionary entries (Heid, 1994; Simons, 1998). Confronted with huge amounts of data, researchers need statistical and computational methods to query it in meaningful ways. Such mastery has been demonstrated by Francis and Kučera (1982) in analysing the 1 million Brown Corpus of American English. They calculated the frequency lists of different word forms and the coefficient of their usage. A similar 1 million word-corpus was built at the University of Lancaster called The Lancaster-Oslo/Bergen Corpus (or the LOB corpus). It had a similar structure to the Brown Corpus but comprised British English (Johansson and Hofland, 1989). Johansson and Hofland did a study of the word frequencies on this corpus to determine the most frequent words. Frequency of usage is crucial to lemmatisation since it guides the lexicographer in determining a headword list. Research on the BNC (Leech et al., 2001) has been attempted involving sophisticated statistics to rank frequency lists of grammatical word classes of the whole corpus, spoken versus written text, and determining distinctiveness of the grammatical word classes of spoken versus written text. Rayson et al. (2002) have analysed the relationship between part of speech frequencies and text typology in the BNC. Levin et al. have used the BNC extensively to demonstrate the role corpus data has in lexical research and the development of a theory that explains and predicts word behaviour. Their research explored the verbs of sound. Other researchers have attempted to assess methodologies of determining which words are particularly characteristic of a text. Kilgarriff (1996) used the BNC to compare the chi-square test, Mann-Whitney ranks test, the t-test, Mutual Information statistic (Church and Hanks, 1989), log-likelihood (Dunning, 1993), poisson mixtures, adjusted frequencies, content analysis (Wilson and Rayson, 1993) and Biber's (1988, 1995) Multi-dimentional analysis in determining which statistical approaches are best suited to identifying words that are characteristic of a text. In the development of this thesis we will explore different statistical approaches to measure similarities and differences in corpus components.

These statistical and computational advancements of querying a corpus are characteristic of developments in research in the English language. Such studies have not been attempted in Setswana.

4.7 Conclusion

In this chapter an attempt has been made to show that, while corpus research stands as one of the most useful approaches to language research, particularly lexicography, in that it can speedily offer information for addressing language related issues and problems, a critical look at the process of corpus construction would help us determine if generalisations drawn from its results should be trusted as true reflection of language use. While corpus linguists are fairly in agreement about the inclusion of language varieties in a corpus, there is still a lack of clarity concerning whether a language population can be known and sampled in all of its varieties. In sampling such varieties, it is not clear how much of each variety is to be sampled. However this has not restricted lexical research to argue that "Corpora like the BNC are designed to provide sample data from which to infer generalisations about the language as a whole, or about particular broad categories of texts…" (Aston, 2001: 75). There are still differences on what it means for a corpus to be balanced and representative of a language from which it was abstracted.

The lack of spoken language and language varieties in many corpora stands as their greatest limitations. This is because the recording and transcription of spoken language is expensive and time-consuming. Communities such as the ones found in many African states face unique challenges to corpus compilation in that their languages are not used in various domains such as: academic writing, media, government and official communication, making text in these domains almost impossible to find. Since automatic transcription is as yet an unsolved problem, it means that attempts of building large corpora of spoken language may remain impossible for some time. The kind of corpus that compilers end up with is therefore the one characterised by Kilgarriff as

> …a corpus which will never be beyond challenge at a theoretical level, but which does nevertheless allow us to address with a degree of objectivity some central questions about the language, where before we could only speculate Kilgarriff (1997: 137).

We have also looked at two corpora, the Brown and the British National Corpus; the former with only a million words, and the later with 100 million words. The two corpora were built about 30 years apart; the Brown Corpus in the 60s and BNC in the 90s. We have inspected their internal structure and revealed that both corpora include samples from different domains to attempt a balanced representativeness of language as used. Both corpora were revolutionary for their times. The Brown Corpus was compiled at the time when hostility was high against impericism, while the BNC is unique for its size and variability. It is through building and querying balanced corpora (Kennedy, 1998; Ooi, 1998: 29) such as the two corpora through advanced statistical and computational approaches that a detailed analysis of a language could be achieved.

Chapter 5

The Setswana corpus compilation

5.1 Introduction

This chapter details the design and compilation of the Setswana corpus used in this thesis. Beyond the thesis the corpus is a resource for corpus investigation of different aspects of the Setswana language research such as morphology, syntax and further investigations of text type variability.

Many linguists and lexicographers look to corpora for linguistic evidence (Al-Sulaiti, 2004). How such corpora are compiled is not always clear and corpus compilers adopt different approaches to compilation (cf. Prinsloo and De Schryver, 2001a and Burnard, 1995). The BNC for instance is considered "a finite, balanced, sampled corpus" (Leech et al., 2001: 1) while the Bank of English is a large organic corpus that is increasingly growing.[15] The varying approaches to compilation have been termed by Church and Mercer (1993: 14) as "a trade-off between quality and quantity". The BNC compilers and those of other balanced corpora are concerned with the quality of the corpus in terms of its constituents, while the Bank of English and industrial laboratories like IBM and AT&T and those compiling organic corpora favour sheer quantity over design niceties. In Chapter 4 we have discussed matters relating to corpus design, amongst these balance and representativeness, corpus mark-up and the representation of speech.

In this chapter we discuss the design and compilation process of the Setswana corpus by showing how the various components of the corpus were collected and compiled and in the case of spoken language how it was transcribed. We also quantify the

[15] http://titania.cobuild.collins.co.uk/boe_info.html

different subcomponents of the corpora in terms of types and tokens, type/token ratio (TTR) and standardized type/token ratio (STTR). Finally, we outline challenges confronted in the compilation of the corpus.

The larger part of the spoken corpus was compiled over a five month period which included fieldwork in Botswana between September and December 2004. The written part of the corpus was collected over a 12 month period. The aim was the collection of as many Setswana language varieties as possible. We use "varieties" as a general term to refer to dialectal varieties, textual types and genres (see Chapter 1, Section 1.3, and Chapter 2 which documents the Setswana text types).

5.2 The design strategy

Setswana is spoken in different dialects by different Batswana tribes and largely in the North-West South African province (see Chapter 2 for a detailed discussion). We originally aimed to sample all or at least most of the Setswana varieties. We planned to record conversations of at least 28 adults (14 male and 14 female aged over 15 years of age). Subjects were to be drawn from both sexes and different age groups of the following Batswana speaking tribes: Bangwato, Bangwaketse, Bakgatla, Bakwena, and Balete. Each subject was to record six hours dialogues. We had also intended to sample University of Botswana students' speech since it was hoped that it would display educated speech with speakers mixing Setswana and English. Our aim was therefore to collect 168 hours of audio-recording. In our funding proposal we had asked for three research assistants who would transcribe audio files for thirty days.

In our compilation of the Setswana corpus we attempted to mirror the BNC methodology (Burnard, 1995). However Setswana presents challenges which are unique to the sociolinguistics of the language which the BNC compilers did not have to contend with. Setswana is used in restricted areas and never or rarely used in other contexts. For instance, the laws and legal proceedings in magistrate's courts and the high court are conducted in English and hardly any written text in Setswana exists. Setswana is not used in this domain save translated speech. The traditional courts (*makgotla*), found predominantly in rural areas, are the ones which use Setswana.

Setswana is also in contact with English and has a historical linguistic contact with Afrikaans. Many speakers are bilingual. They therefore mix English and Setswana (see Section 4.4.3). Code-switching, code-mixing and diglossia and the bilingualism and multilingualism of the speakers compounded the problem of spoken language transcription. Additionally, the practical considerations of time and funding meant that the text had to be scaled down to a size that was manageable for the PhD research.

5.3 Overall corpus statistics

We begin by presenting the overall statistics of the corpus and of the broad subcorpus portions of spoken and written language. We then proceed to looking in considerable detail at the design and compilation of both the spoken and written language corpus sections.

The total Setswana corpus compiled is over 13½ million tokens, 13,695,965 tokens to be exact (for a discussion of tokens and types see Chapter 3). Ninety four percent of the corpus is the written component while the spoken component is 6%. Table 21 gives the sizes of the tokens, and the type/token ratio (TTR) and standardized type/token ratio (STTR) measures of the broad components of the written and spoken parts of the whole corpus.

Table 21: Overall corpus statistics

File size (bytes)	95,009,785
tokens	13,6975,965
types	372,513
type/token ratio (TTR)	2.83
standardised TTR (STTR)	33.58

The type/token ration (TTR) is calculated by dividing types by tokens and multiplying by 100. By types we refer to the *different types* of words that occur in a document while by tokens we refer to the count of every word regardless of its repetition. Thus if the word *gore* occurs in a document 75 times, it is said to constitutes a single type but 75 tokens.

The TTR however varies widely in accordance with the length of a text; with shorter texts, the statistic is much more likely to give higher TTR, while longer texts result with a smaller TTR (Malvern and Richards, 2002). Becasuse of this phenomenon McKee et al. show that TTR measures are flawed,

> …because the values obtained are related to the number of words in the sample... samples containing larger numbers of tokens give lower values for TTR and vice versa. …as longer and longer samples of language are produced, more and more of the active vocabulary is likely to be included and the available pool of *new* word types that can be introduced steadily diminishes. ...ıt is also the case that however small the sample is, as more and more tokens are taken, the likelihood is that (because of repetition of previously included types) the cumulative number of types will increase at a slower rate than the number of tokens and the TTR values will inevitably fall (McKee et al., 2000: 323).

The solution to this challenge is to compare equal sized text types. The results of comparing Setswana texts would have been much more significant if the text types were of the same size such as in the LOB and Brown Corpus subcorpora. A more reliable measurement is that of the standardized type/token ratio (STTR). We use Wordlist tool of WordSmith Tools to run the measures. The ratio for STTR is calculated at every specified number of tokens and an average of the different ratios computed. STTR is computed every **n** words as Wordlist goes through each text file. For the experiments, n = 1,000. In other words the ratio is calculated for the first 1,000 running tokens, and then calculated afresh for the next 1,000, and so on to the end of the text or corpus. A running average is computed, which means that we get an average type/token ratio based on consecutive 1,000-word chunks of text. Texts with less than 1,000 words get a standardized type/token ratio of 0. STTR measures are attractive since they can compare type/token ratios across texts of differing lengths since what they do is segment a corpus into comparable chunks and calculate the type/token ratio for each. In Section 6.5 we use STTR measures to compare corpus chunks.

Another way of looking at the whole corpus is through frequency profiling. Table 22 gives the statistics of the top 20 tokens in the whole corpus.

Table 22: Top 20 Setswana tokens[16]

Rank	Word	Freq.	%
1	a	686,492	5.01
2	go	418,088	3.05
3	e	413,176	3.02
4	le	358,736	2.62
5	o	336,417	2.46
6	ba	315,243	2.30
7	ka	290,557	2.12
8	ke	242,497	1.77
9	ya	228,511	1.67
10	mo	193,181	1.41
11	re	158,644	1.16
12	ga	149,529	1.09
13	fa	143,385	1.05
14	se	132,649	0.97
15	gore	125,686	0.92
16	di	124,651	0.91
17	ne	97,129	0.71
18	wa	94,822	0.69
19	tsa	92,885	0.68
20	sa	81,099	0.59
TOTAL		**4,683,377**	**34,2**

The most frequent token is *a* with a frequency of 686,492 which is about 5% of the whole corpus. Within the top 20 ranked tokens, the word frequency has declined to 81,099 which is about half a percentage (0.59). It is also clear that the most frequent tokens constitute a large percentage of the corpus. The most frequent 20 tokens constitute just over 34% of the whole corpus (over 4½ million tokens). As Table 23 shows, close to 55% of the whole corpus (over 8 million tokens) is made up by the top 1000 tokens and the top 10 tokens in the corpus constitute over 25% of the whole corpus. Comparatively, the Brown Corpus' 10 most frequent tokens account for 23% of the whole corpus (Baroni, 2006: 5).

[16] In this thesis since we count tokens as graphical units, homographs are counted as single tokens in all tables.

Table 23: Top 1000 token-ranges and percentages in the whole Setswana corpus

Range	tokens	%
1-10	3,482,898	25.43
11-20	1,200,479	8.77
21-30	556,051	4,07
31-40	295,948	2.17
41-50	205,738	2
51-100	595,259	4.34
101-200	632,192	4.59
201-300	347,476	2.49
301-400	249,737	2
401-500	191,595	1.23
501-1000	592,026	1.3
TOTAL	**8,349,399**	**54,32**

5.4 The Zipfian distribution

This rapid decline in frequency with few words having very high frequencies is common in corpora and has been used as a reason why large corpora are needed to accurately account for low frequency words (Fillmore et al., 1998). The rapid frequency decline in corpora has been explained by the famous Zipf's law. Zipf (1949) was concerned with such quantitative analysis such as the relationship between the frequency of words in text and text length, 'the frequency of words and their antiquity' (Kennedy, 1998: 10) and the relationship between the rank order of an item in a word frequency list and the number of occurrences or tokens of that item in a text. Zipf's law has been defined formally by Evert and Baroni (2005: 2/3) as follows: the *frequency* f_n of the *a* word type *w* is inversely proportional to its Zipf rank *n*, i.e., the rank of *w* in a list of all word types ordered by decreasing frequency. Zipf's law therefore holds that the relationship between the frequency of use of a word in a text and the rank order of that word in a frequency list is a constant (f.r=c) (Kennedy, 1998: 10). "Consequently, a very small number of words occur extremely often, and a very large number of words occur very infrequently" (Atkins et al., 2001: 53; emphasis that of the authors).

In the discussion of Zipf's law Gomez (2002: 235) shows that Zipf was one of the first linguists to prove the existence of statistical regularities in language with his best known law which proposes a constant relationship between the rank of a word in a frequency list and the frequency with which it is used in a text. This is because the

relationship between *rank* and *frequency* is inversely proportional. In addition, Zipf thought that the *constants* are obtained regardless of subject matter, author or any other linguistic variable.

On Zipfian distribution, Kilgarriff notes that

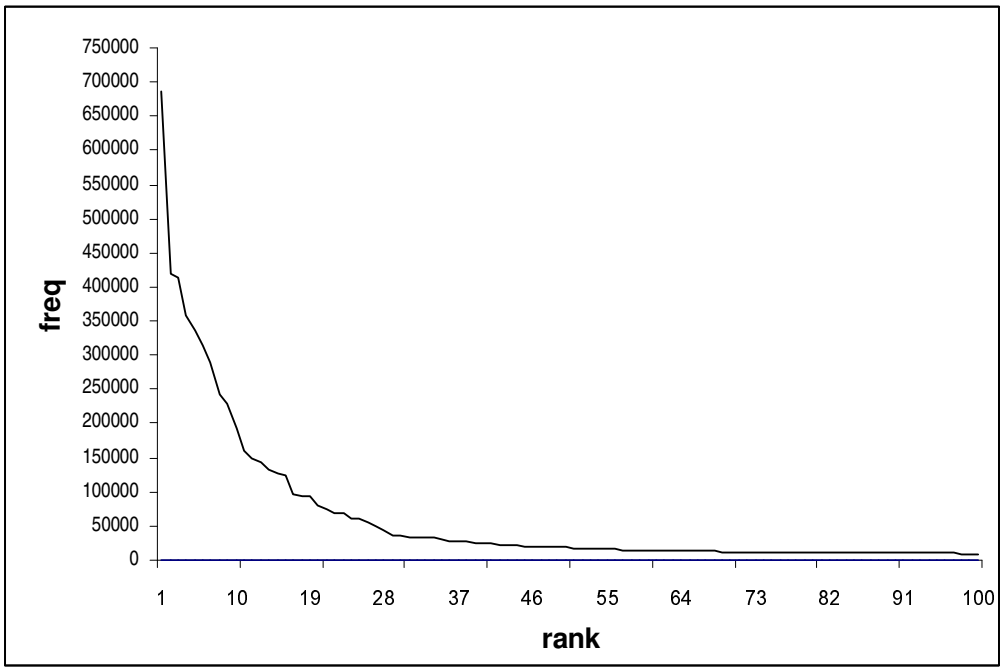
> In a Zipfian distribution, the most common item has twice as many occurrences as the second most common, three times as many as the third, a hundred times as many as the hundredth, a thousand times as many as the thousandth, and a million times as many as the millionth (Kilgarriff, 1996: 2).

Baroni (2006) notes that Zipf (1949, 1965) has observed that frequency is a non-linearly decreasing function of rank (decreasing more sharply among high ranks than among low ranks), and proposed the following model, which became known as Zipf's law, to predict the frequency of a word given its rank:

$$f(w) = \frac{C}{r(w)a}$$

In the formula, *f(w)* and *r(w)* stand for frequency and rank of word *w*, respectively. *C* and *a* are constants to be determined on the basis of the available data. To understand why this is a plausible model, assume for now that *a* = 1, so that the equation can be simplified to *f(w) = C/r(w)* . Then, the most frequent word in the corpus, having rank 1, must have frequency *C*. In our corpus the most frequent word, *a*, has frequency 686,492 and thus we set C = 686,492. According to the formula the second most frequent word is predicted to have frequency 686,492/2 = 343,246, which is half the frequency of the first word. The third most frequent 686,492/3 = 228,831. Baroni (2006: 11) points out that the model predicts a very rapid decrease in frequency among the most frequent words, which becomes slower as the rank grows, leaving very long tails of words with similar low frequencies. This is true for the Setswana language as we see in the graph below of the most frequent 100 tokens.

Figure 5: A rapid frequency decline in the top 100 words



The Zipf's law has been offered by Manning and Schutzë (1999: 24) as:

There is a constant k such that $f \cdot r = k$

In Table 24 we empirically evaluate Zipf's law with the top 20 Setswana tokens from the corpus.

Table 24: Top 20 Setswana tokens

Word	**Freq.** (*f*)	**Rank** (*r*)	$f \cdot r$
A	686,492	1	686,492
Go	418,088	2	836,176
E	413,176	3	1239,528
Le	358,736	4	1434,944
O	336,417	5	1682,085
Ba	315,243	6	1891,458
Ka	290,557	7	2033,899
Ke	242,497	8	1939,976
Ya	228,511	9	2056,599
Mo	193,181	10	1931,810
Re	158,644	11	1745,084
Ga	149,529	12	1794,348
Fa	143,385	13	1863,992
Se	132,649	14	1857,086

Gore	125,686	15	1885,290
Di	124,651	16	1994,416
Ne	97,129	17	1651,193
Wa	94,822	18	1706,796
Tsa	92,885	19	1764,815
Sa	81,099	20	1621.980

Our Table 24 results can be summarised in a similar manner as those of Manning and Schutzë who observe that while Zipf's law holds for parts of the list; it is off for the very top tokens on the list.

The discussion of the Zipfian distribution on this chapter is significant since it has a bearing on word counting which is at the centre of experimentation in this thesis. Manning and Schütze, however caution that a Zipfian distribution is better perceived as a rough estimate of how frequencies are distributed and not as a law (Manning and Schütze, 1999: 24).

5.5 Corpus components

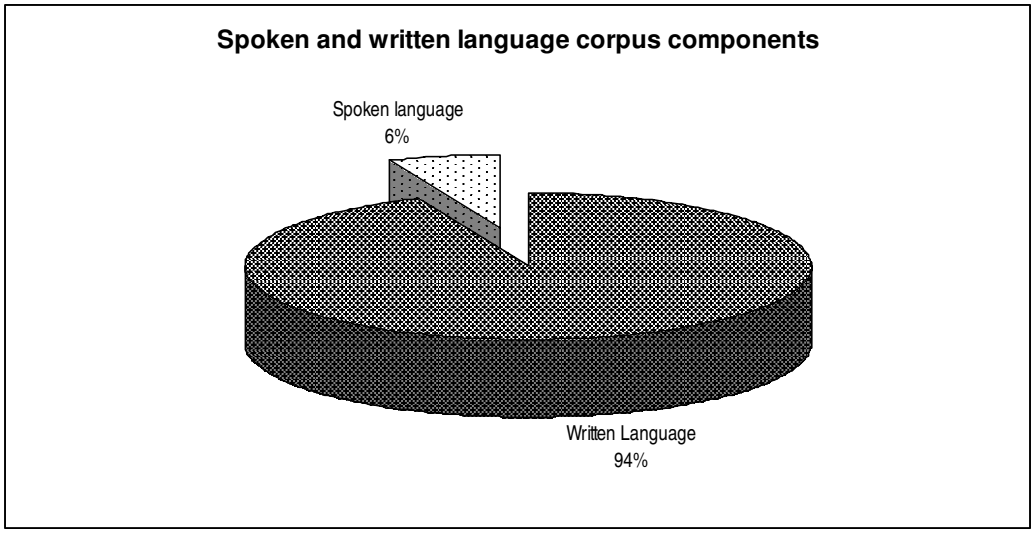
Below we give the different components of the whole corpus on the basis of tokens, types, TTR and STTR. First, we calculate the large corpus components of written and spoken language. The results are given in Table 25.

Table 25: The corpus written and spoken components

Text type	Tokens	Types	TTR	STTR
Written language	12,831,759	358,182	2.90	33.63
Spoken language	840,400	38,118	4.54	32.94

Table 25 reveals the corpus components divisions with the bulk of the corpus being material from the written language. While there are huge numerical differences between spoken and written language, both in terms of tokens and types, the differences on the basis of STTR between the two are minor (33.63 for the written language and 32.94 for spoken language). Figure 6 demonstrates that 94% of the corpus is written language material while 6% is spoken language.

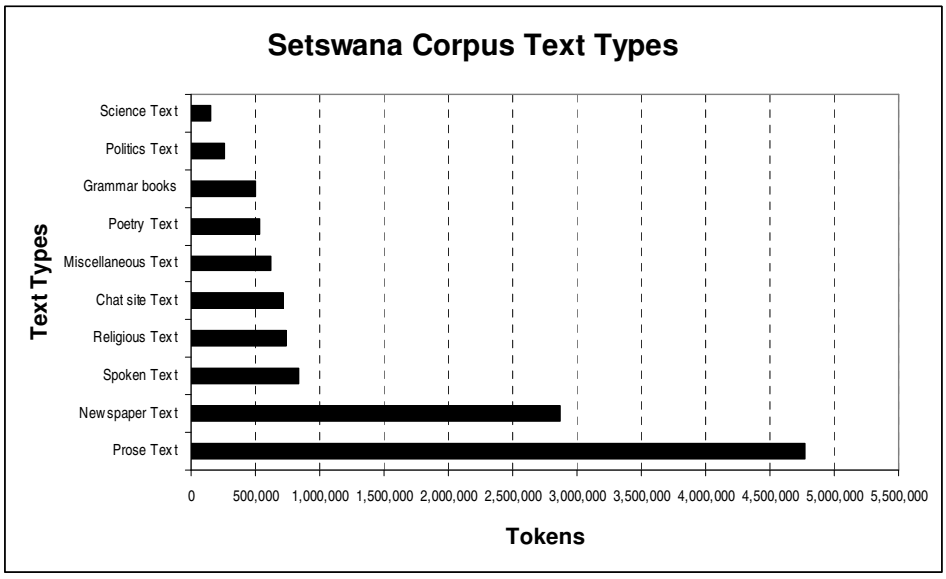
Figure 6: Spoken and written language corpus components pie chart



5.5.1 Text types in the corpus

Having looked at the broad sections of spoken and written language, we turn our attention to the text types in the corpus. In Figure 7 we plot the corpus types on a graph and what becomes apparent immediately is that Prose text occupies the largest portion of the corpus, followed by Newspaper text and Spoken text. Science text has the fewest tokens.

Figure 7: Setswana corpus text types



We now look at the different components of the corpus in detail. First we consider Spoken language components and then the written language components.

5.5.2 The spoken language components

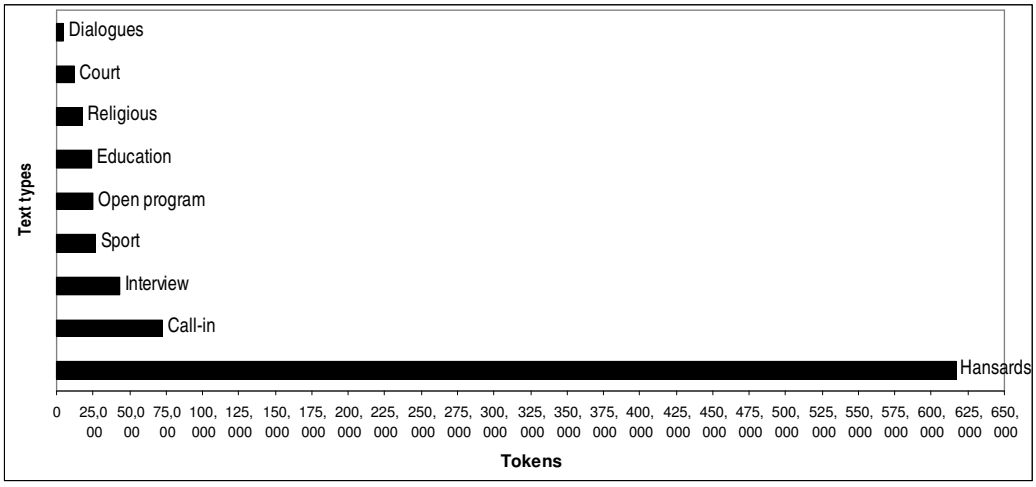
The overall spoken component of the Setswana corpus has 840,400 tokens and 38,118 types. It has a type/token ratio of 4.54 and the STTR of 32.94. Table 26 shows the breakdown parts of the spoken component.

Table 26: Spoken components statistics

Text-type	Tokens	Types	TTR	STTR	%
Hansards	616,695	33,581	5.45	35.51	73
Call-in	72,634	4,264	5.87	27.05	8.63
Interview	42,882	3,795	8.85	26.66	5
Sport	26,618	2,162	8.12	30.12	3.16
Open program	25,194	3,968	15.75	35.16	3
Education	23,545	1,329	5.64	25.20	2.80
Religious	17,736	2,210	12.46	29.14	2.11
Court	12,216	1,829	14.97	34.59	1.45
Dialogues	4,207	599	14.24	25.07	0.50

Table 26 also shows that 73% of Spoken text is text from Hansard and the remaining 27% is shared between eight other sources (see Figure 8). All spoken language is spontaneous speech and not scripted.

Figure 8: Spoken components statistics



While high levels of Hansard material in the corpora may appear to distort the spoken material, Hansard text is attractive in that it is text on a variety of subjects handled in Parliament and has a potential of contributing a variety of types. Its diversity is in part supported by a high STTR of 35.51.

5.5.3 The written language components

The written component of the corpus occupies the largest part of the corpus at 94%. It comprises about 12,831,795 tokens, 358,182 types, with a STTR of 33.63. Table 27 reveals that Prose has the largest number of tokens and types, followed by Newspaper text. Science text has the smallest number of tokens. Although Science text has the smallest number of tokens, it is Politics that has the smallest vocabulary with the lowest number of types.

Table 27: Overall statistics of the written subcorpus

Text Types	Tokens	Types	TTR	STTR
Prose Text	4,772,704	289,270	6.00	38.55
Newspaper Text	2,870,300	74,497	2.60	27.20
Religious Text	735,061	30,539	4.15	34.87
Chat-site Text	712,445	37,403	5.26	44.89
Miscellaneous Text	616,181	49,725	8.07	34.30
Poetry Text	530,261	47,235	8.91	43.43
Grammar books	504,559	35,386	7.01	37.05
Politics Text	262,652	10,782	4.11	30.23
Science Text	154,398	10,878	6.87	33.30

Table 28 on the other hand, presents the results ordered on the basis of STTR. The STTR measures are ordered in decreasing frequency. The evidence reveals that Chat-site text has the largest lexical density. This is to be expected since Chat-site text has high levels of code-mixing, code-switching, unconventional spelling patterns, and cover diverse topics which lead to high levels of STTR. Newspaper text has the smallest STTR.

Table 28: STTR measures of the written subcorpus

Text Types	STTR
Chat-site Text	44.89
Poetry Text	43.43
Prose Text	38.55
Religious Text	34.87
Miscellaneous Text	34.30
Spoken Text	33.86
Science Text	33.30
Politics Text	30.23
Newspaper Text	27.20

Poetry is generally believed to use "rich language" characterised by proverbs and a variety of figures of speech. This appears to gain support from the high STTR numbers. For newspaper text to have the lowest STTR may be a result of the use of simple language to achieve communicative efficacy by a newspaper.

5.5.4 Newspaper text breakdown

Newspaper text is however complex since it comprises news, sport, letters to the editor, editorials, columns and other sections. To study these different components we have divided newspaper text into further sub-divisions of News, Arts & Culture, Sports, Business and Letters. With such subdivisions, we are able to study such specialised areas of newspaper reporting such as Sports and Business in considerable detail over and above looking at newspaper text as a unit.

Figure 9: Newspaper text division

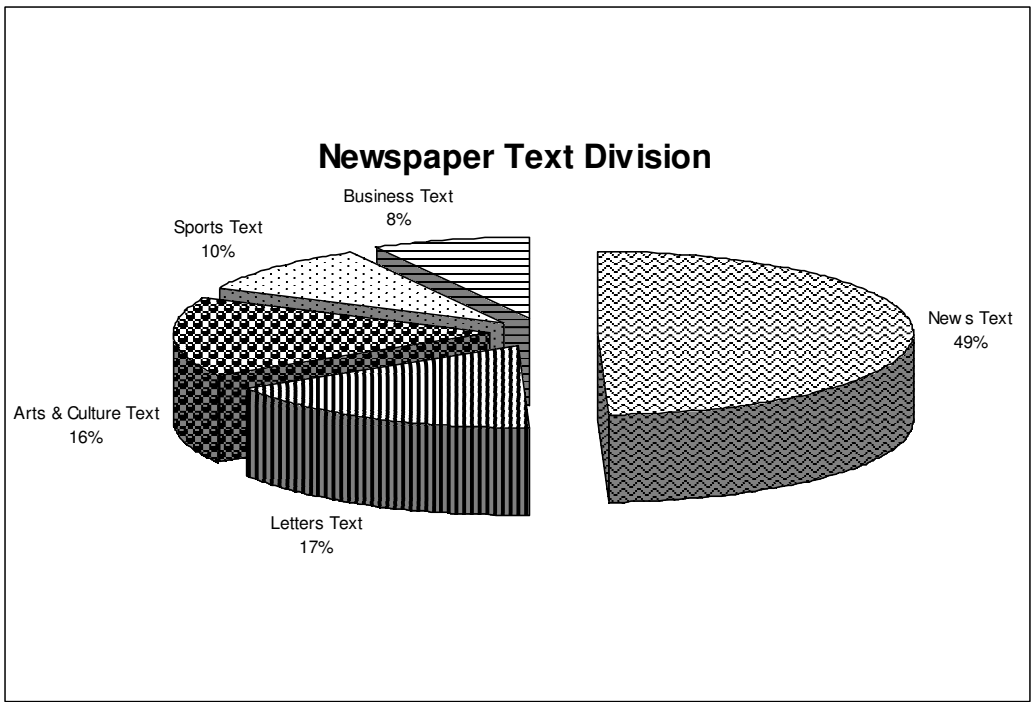


Figure 9 reveals that about 50% of the newspaper text is News while about another 30% is made up of Arts and Culture and Letters, 16% and 17% respectively. The Letters subcorpus comprises letters to the editor, editorials and newspaper columns. Business text has the smallest percentage of 8%. Business text comprises adverts and business news. Table 29 shows that News text has the highest STTR (29.75) followed by Arts and Culture text (26.54), Sports text (25.79), Letters text (22.73) while Business text (20.6) which has the lowest number of tokens, has the lowest STTR also.

Table 29: Newspaper component statistics

Text Types	Tokens	Types	TTR	STTR
News Text	1,415,836	57,084	4.03	29.75
Letters Text	490,933	18,031	3.67	22.73
Arts & Culture Text	455,418	19,157	4.21	26.54
Sports Text	274,764	11,339	4.13	25.79
Business Text	233,349	9,780	4.19	20.60

The different components of the Newspaper text in Table 29 will be compared against other parts of the corpus in Chapter 6 and Chapter 7.

5.5.5 Prose text breakdown

The Prose section of the corpus has the largest number of tokens partly because of the large number of published Setswana novels included in this section. However Prose does not only include novels. Included in this section are folklores/folktales, collections of short stories, children's literature, cultural texts such an anthology of proverbs, sayings and riddles. Included also are cultural books about chieftaincy and the Setswana culture and language in general. The texts also comprise some online documents such as student tests and some online Setswana newsletters. The summary of this information is in Table 30.

Table 30: Prose component statistics

Text types	Tokens	Types	TTR	STTR
Novels	3.787.585	212.543	6	32.13
Cultural texts	662.756	29.090	4	32.70
Language texts	291.486	19.360	7	33.52
Magazine	78.990	7.531	10	32.13
Online texts	9.284	1.594	19	32.02

The novels have the largest number of tokens while online texts have the largest number of TTR. A large TTR is characteristic of smaller tokens as evident with online texts while large tokens, as in novels and cultural texts, are largely characterised by smaller TTR. STTR across all texts is fairly similar suggesting that all prose texts are similar.

5.6 The compilation of corpus components

Ideally it would be attractive for linguists to analyse all communication acts ever uttered by members of a language community and all written material ever produced. However such a phenomenon is beyond reach of linguists first because the data would be far too large to explore in its entirety and second, because the quantity of all utterances of speakers and that of all written material is unknown. Teubert (2001:

129) also argues that it "is the responsibility of the linguist to limit the scope of the universe of discourse in such a way that it may be reduced to a manageable corpus, by means of parameters such as language (sociolect, terminology, jargon), time, region, situation, external and internal textual characteristics, to mention just a few."

Next, we discuss how we have limited the scope of our text selection to be included in the corpus. We examine the compilation of the broad corpus components of spoken and written language and present measurements of their various sub-components. First, we consider the spoken language and then proceed to the written language component.

5.6.1 Spoken language component compilation

i. Sampling

As with any sampling, some compromise had to be achieved between what was "theoretically desirable and what was feasible" (Burnard, 1995: 21) for our research. With five months within which to begin and complete corpus compilation, with three research assistants and two computers, there were limits to what could be achieved. Indeed our approach to the compilation of the spoken text is characterised accurately by Atkins et al.

> The difficulty and high cost of recording and transcribing natural speech events lead the corpus linguist to adopt a more open strategy in collecting spoken language (Atkins et al., 1992: 3).

It is the high cost of recording and transcribing natural speech events which led us to a different approach in text collection and transcription which will be discussed briefly later.

In recognition of our limited time and resources, the corpus compilation had to be scaled down to an achievable size which was nonetheless large enough to be queried for linguistically interesting data.

ii. Recording

The corpus contains recordings of sermons, family dialogues, funeral services, classroom interactions, radio and television debates, court transcriptions and other spoken text, recorded using micro-cassette tape recorders. Conversations, speeches and other dialogues were recorded as unobtrusively as possible ensuring that the material gathered was as natural and as spontaneous as possible. For instance in classroom recordings, teachers were trained on how to record themselves and were given tape recorders to take to class. The researcher avoided going into a classroom to record a teacher since it was felt that this could create tension and make the teacher feel under observation which could lead them to modify their speech. In other cases a different approach was used. For instance in funeral recordings, permission was sought from the family in advance and different speech makers in the service/ceremony.

The Department of Information and Broadcasting also gave us access to recordings of radio and television programs like call-in programs and live radio debates. These recordings are attractive since they feature different speakers of Setswana dialects and are on a variety of topics.

Below we briefly discuss different recordings and categorise them in terms of BNC labels.

- **Educational and informative**

  Classroom interaction: Classroom interactions were recorded in different schools. Since the Setswana language is only used in the teaching of the Setswana language and literature in secondary schools, the recordings capture only Setswana lessons. In the recordings, since it is the teacher who carried the tape recorder, what has been recorded is the teacher's voice while the students' voices are virtually inaudible.

- **Public or institutional**

  Sermons: Sermons were recorded in different church denominations and funeral services. Recordings were done in Gaborone and Kanye.

  Parliamentary proceedings: Parts of the Botswana Hansard in Setswana were scanned. Most debates in parliament are in English, however members sometimes use Setswana. We therefore looked for Setswana chunks in the Hansards and scanned them for inclusion in the corpus.

  Radio debates between candidates for parliamentary seats were also recorded and transcribed for inclusion in the corpus.

  Legal proceedings: We were fortunate to have access to transcribed legal proceedings used by Thekiso (2001) for PhD research on court discourse in Botswana. These were incorporated in the corpus to represent legal text.

  Funeral services: We attended and recorded three funeral services. Funeral services in Botswana are usually characterised by short speeches from various people who may include a village elder, a nurse, a councillor, a representative of a burial society, a pastor, and many others. These were recorded, transcribed and the text included in the corpus

- **Leisure**

  Broadcast chat shows and phone-ins: Unscripted chat shows and phone-ins on different subjects were recorded from Radio Botswana and Botswana Television.

  Sports commentaries: Only football commentaries were recorded and transcribed.

iii. Transcription

Our transcription scheme was developed following Crowdy (1994: 25) who suggests that "[t]he design of any transcription scheme should involve considerations of: who is the transcription for? How will it be used? What are the important features?" The Setswana corpus was primarily compiled to aid thesis research in comparing corpora segments for lexicography. Wordlists of its different parts have been generated and compared against other lists drawn from other parts of the corpus (see Chapter 6 and Chapter 7). The corpus will also be accessed for particular linguistic features or viewed in concordance form. It is envisaged that beyond current research, the corpus will prove useful as a national resource and may be of interest to discourse researchers, grammarians and general linguists.

Non-linguistic factors, like time and budget, have impacted on the transcription design. The corpus does not mark any paralinguistic phenomena like whispers, laughs, and coughs. Non-verbal and non-vocal events for example animal noises or passing lorries are also not marked. There are also no markings for significant pauses within or between utterances. All these paralinguistic phenomena, though significant in other studies, were deemed not critical for our experiments. The transcription principle we adopt is simple with limited mark-up.

Plays text has been marked-up because of the unique challenges it poses. The challenge is raised by the repeated personal names. Plays comprise sequences of personal names followed by a character's words which need to be treated as metadata that they do not interfere with the frequency counts. We marked speakers' names as meta-text and marked them up in such a way as to exclude them from the counts. An illustration follows:

<c>Bothata</c> A re o lomiwa ke eng?

<c>Thekiso</c> Ke ka bo ke akga loleme fa ke ka go raya ka re ke itse se se mo jang.

<c>Bothata</c> Tlhokomologa tseo ngwanaka a re robale. Gongwe o itse se a se lwelang, o tlaa itlhalosa fa a na le kgang. Tshu! ke šele jang. Letsatsi le sala le tlhola le kgwisa kolobe diphulo. Tima lebone

foo mma.

Since we envisaged investigating code-switching and code-mixing in this study English words in the spoken part of the corpus were marked up. For example

1. Go ya ka <eng>Assistant Superintendent</eng> Mmoloki:
"According to Assistant Superintendent Mmoloki".

2. Go lebiwa <eng>next</eng> <eng>structure</eng> ya <eng>society</eng>
"The next structure of the society is considered."

There were also challenges with spoken language transcription. Early in the transcription stage it became clear that the assistants had problems with Setswana word divisions. This problem is common amongst Batswana and it is a result of poor literacy in the Setswana language beyond secondary education. The errors they displayed included, amongst others: [*are*] instead of [*a re*] "he/she said", or confusing [*ene*] "him/her" with [*e ne*] "it was". Other problems concerned failure to identify sentence boundaries in speech. Because of these problems post-transcription and editing were undertaken by the author.

5.6.2 Compiling the written language component

i. Sampling

The scope of Setswana texts is limited. Most Setswana texts are published for the school curriculum. The majority of them are therefore grammar books and literature material (novels, plays and poetry books) for Setswana classes at both primary and secondary school levels. The texts are limited to materials for language and literature classes. Other subjects like Mathematics, Science, Agriculture and Art are taught in English, and therefore use texts written in English. Material in such subjects could therefore not be included in the corpus. Hardly ever do people read leisure texts in Setswana partly because these are rare and partly because there is no literacy culture in the Setswana language, beyond secondary school education. School and public libraries and bookshops have small numbers of Setswana books. There are neither

bestsellers lists nor literary prizes which could be inspected for potential texts inclusion. Most included novels, plays and poetry had either been in the curriculum or were currently used in schools. All the texts were published after 1980. This date was not an intentional cut-off date, texts in Setswana published before 1980 are hard to find. The general rarity of texts, and their small size (in terms of number of words), necessitated the inclusion of whole texts in the corpus.

The corpus includes texts from two newspapers: *Mokgosi* and *Naledi*. *Naledi* is an insert in the largest private daily, *Mmegi*, while *Mokgosi* was the only weekly newspaper that wrote exclusively in Setswana. The *Mokgosi* newspaper closed down in 2005. The Newspaper text is divided into five broad categories: Arts and Culture, Business, Letters (letters to the editor and columns), News and Sport.

The Setswana corpus also contains miscellaneous texts including student essays and letters from junior secondary schools, and the complete text of the national vision. There is also religious text (Christian, Bahai, Islamic texts). There are also political texts, Science text, Business text (e.g. from Botswana Meat Commission and Botswana Telecommunication Authority). Magazines in Setswana are hard to find. However, the *Kutlwano* magazine, which is predominantly written in English, has stories in Setswana which we were able to include in the corpus.

The corpus also includes Web text. In collaboration with Kevin Scannell, of St Loius University (USA), we mined the Web for Setswana text using *An Crúbadán*, a Web crawler for the "automatic development of large text corpora for minority languages" (Scannell, 2007)[17]. From this automatic mining of the Web we were able to build approximately half a million words. This part of the corpus together with another one million words from Macmillan has been used to build the first Setswana spellchecker (*aspell-tn*, *ispell-tn*, *myspell-tn*) used by OpenOffice[18]. The mined texts include different kinds of documents including religious literature, law, outlines of different government projects, health literature, examination question papers and other educational material and different kinds of literature. These files were added to their

[17] http://borel.slu.edu/crubadan/index.html

[18] http://lingucomponent.openoffice.org/spell_dic.html

appropriate text types in the broad Setswana corpus. However the crawler did not download certain linguistically interesting files from the Web, specifically message boards.

We downloaded Edumela web-pages of message board text[19]. Edumela is a chat-site used mainly by Batswana students studying at colleges and universities in and outside Botswana. The language used on the chat boards is relaxed, colloquial and is characterised by code-switching, code-mixing and greater levels of English use, especially in discussions on science and technology. The inclusion of chat-site documents in corpus compilation finds support in Villasenor-Pinedar et al. who argue that it is closer to naturally occurring speech. They argue that:

> Because many people around the world contribute to create the Web documents, most of them have informal contents, and include many everyday as well as non-grammatical expressions used in spoken language. This situation allows [for] …the construction of very large corpora combining good written grammatical text and free text closer to the spoken language (Villasenor-Pinedar et al., 2003: 393).

Villasenor-Pineda et al.'s observations concerning the Web message board text is accurate concerning Edumela text since the text resembles that of colloquial Setswana. The following illustration from Edumela show that Edumela text is complex, comprising colloquial language, English and Setswana. The text is largely written in English with colloquial words italicised while formal Setswana words are bolded. English translations of both colloquial and formal Setswana are in brackets.

> **Owaaii** *girlie* **tota** (Uh! Girl, truly) there is nothing we can do for you except to advise you **gore** (that) try to leave that man *coz* **le wena** (because you too) at least you know **gore** (that) he is using you. **Jaanong ha o re** (Now if you say) you don't want him to leave his wife, **mme gape o** (but again you are) jealous of the wife **o raya jang**? (what do you mean?) What do you want? *(Italicising and translation in brackets mine).*

[19] www.edumela.com

The example illustrates the kind of code-mixing which is particularly common amongst the university students and urban dwellers in general. Although this text was written and not spoken, the code-mixing that characterise it is typical of spoken language particularly amongst the young educated and urban Setswana speakers. We have discussed some of this code-mixing and code-switching and colloquialism in Section 4.4.3 of this thesis.

5.6.3 Spoken language ethical matters

Dealing with human subjects in corpus compilation raises ethical matters relating to subjects' confidentiality. It is no wonder Martin and Mauldin (1997: 570) excluded texts from the Creek corpus which they deemed to be of a highly personal nature; those that criticised other community members or included personal names. For our purposes, participants recorded at schools, homes, churches and funerals were guaranteed confidentiality and anonymity. Personal names, addresses, phone numbers and car plate numbers have been removed from the corpus to ensure that participants are not identified. Subjects completed and signed the Participation Consent Form (see Appendix 2) which explains the corpus compilation process and assures them of the protection of their confidential information. The participants also had to complete a Conversation log (see Appendix 3) which details where the dialogues took place (village, town etc), what the subjects were doing during the recording. The conversation log also includes a place where a list of first names of people in the dialogue could be entered. Both the Participation Consent Form and the Conversation log were translated into Setswana for people whose knowledge of English is limited. For the illiterate, the Participation Consent Form was read to them and they had to accept on tape that they agreed to be recorded. In the case of schools, school-heads (headmasters) were sent letters requesting permission to make the recordings (see Appendix 4 and 5). The school-head then met the members of the Setswana department in the school to discuss the research and subsequently offer consent or refuse it.

5.6.4 Written language ethical matters

In the case of written text, publishers were sent letters (Appendix 6) requesting text. Various departments and organisations were visited and permission to have access to Setswana text sort. Permission was either granted, refused or in most cases Setswana text was unavailable.

The government of Botswana requires that individuals conducting research in Botswana should apply for a research license with the Permanent Secretary of the Ministry of Labour and Home Affairs (Botswana) before research begins. Such a license was obtained.

5.7 Conclusion

In this chapter we have mapped out the compilation of the Setswana corpus which we use for experiments in this thesis. It is about 13 million words and covers text from different varieties of Setswana including, novels, plays, newspapers, grammar books, spoken language covering court transcripts, call-in programs, television debates, funeral services, classroom interaction and sermons. The Setswana corpus design and compilation was influenced by both the BNC and the Russian Corpus (Sharroff, 2004).

We have also discussed the recording and transcription process and the ethical issues confronted.

We have also discussed the Zipfian word distribution as it relates to the Setswana corpus and seen that most of the corpus is made up of high frequency words. The most frequent word has been found to be *a* with a frequency of 686,492 which is about 5% of the whole corpus. The most frequent 20 words have been found to constitute over 34% of the whole corpus (over 4½ million tokens). About 55% of the whole corpus is made up by the top 1000 tokens and the top 10 words in the corpus constitute over 25% of the whole corpus. Such a situation necessitates large corpora for the study of particularly low frequency words.

Since this is the largest Setswana corpus with a diverse collection of texts, that we are aware of, it is a significant resource for future Setswana language research in general linguistics and lexicography. The corpus may be used for the development of monolingual and bilingual dictionaries, thesauruses, and grammars. The publications and tools developed from the corpus will benefit mother-tongue language users, researchers, teachers, students and publishers.

The corpus like many corpora has large sections of written language and smaller sections of transcribed spoken material. Ninety four percent of the corpus is the written component while the spoken component occupies 6%.

In the next two chapters (Chapters 6 and 7) we use the corpus in experiments to measure lexical density across a variety of text types.

Chapter 6

Measuring text type diversity

6.1 Introduction

Lexical researchers are interested in comparing different language varieties to measure language variation or describe lexical qualities of a subcorpus (Kilgarriff, 1996; Rayson et al., 1997; Kilgarriff, 2001; Leech et al., 2001; Rayson et al., 2004). For the purpose of this thesis we want to determine whether a corpus with texts from various text types is "better suited" for lexicography than a corpus compiled with texts from a restricted domain. We proceed from the assumption that text variability in corpus compilation is desirable. The assumption, however, demands empirical verification. Such verification can be achieved through experiments which compare corpora and corpus components. To perform such comparisons accurately, statistical methods are employed since we agree with Kilgarriff (2000: 109) that "Lexicographers need the skills and or the software to navigate through sometimes huge numbers of corpus instances." They need a mastery of statistical methods and natural language processing to make sense of the data. In this study the statistical analysis is conducted through the use of WordSmith Tools.

In Section 6.2 we calculate keywords for Science and Technology, Politics, Poetry, Plays, Grammar, Arts and Culture, Religious and Hansard texts and interviews text from spoken language. The top 100 keywords from each genre or text type are presented. For a corpus to represent the general language, it must be designed in such a way that it includes a variety of text types from the language which it represents. Oostdijk has argued that,

> [i]t is a well-known fact that a language is not a homogenous phenomenon but rather a complex of many varieties. The existence of linguistic variation is something linguists have long been aware of (Oostdijk, 1988: 12).

We therefore intend to show through keyword analysis that different text types generate different keywords that are particular to them. Such a result would give support to the argument that a corpus that reflects linguistic variability of a language community must be compiled with a variety of texts drawn from different text types. The aim is to measure if different text types contribute distinct words. If this is found to be the case, then such a finding would prove significant to corpus design for lexicography in general, and particularly to this thesis.

We follow keyword analysis by measuring type/token of various text types at 10,000 tokens intervals. The measurement determines the rate at which types grow at specific points across text types. The measure aims to show that different texts, even with the same number of tokens, contribute distinct types. We also take corpus samples from three text types and combine them together and measure them against the different text types from which their parts were compiled.

We follow the type/token measures experiment by testing how frequency lists from different text types and the frequency lists from the three compilations [POEGRACHAPLA (Poetry, Grammar, Chat-site and Plays), SCIPOLBUSREL (Science, Politics, Business and Religious text) and PROSPONEW (Prose, Sport, Miscellaneous and Newspaper text)] perform when juxtaposed to the frequency list generated from the whole corpus. The purpose of the experiment is to measure how individual lists extracted from various text types compare to a wordlist extracted from the whole corpus.

We conclude the section on wordlist experiments by comparing the most frequent 100 words of the spoken part of the corpus and those of the written part, against those of the most frequent 100 words of the entire corpus. It is hoped that it will be evident that a corpus with only spoken or only written text is inadequate in the isolation of words which could be used for a headword list. Rather an attractive approach is to include both written and spoken material in a corpus.

Additionally, we conclude the chapter by further testing whether text type diversity is crucial to the quality of the words for inclusion in a dictionary. Two 5,000-word list chunks will be compared. The first chunk simulates a wordlist drawn from an opportunistic corpus with its text type limitations since it is derived exclusively from prose text. The prose text is chosen since much of text in many African languages will be of a prose type. Most of such text comprises novels. The most frequent 5,000 words are therefore derived from the prose text. The other 5,000 words are derived from a corpus comprising a variety of text types. The two wordlists are then tested for the presence of terms business, religion and vulgarities.

We begin by looking at keyword analysis.

6.2 Keyword analysis

In this chapter we return to keyword analysis, a subject we introduced in Chapter 3 (Section 3.8).

Our calculations do not make distinctions between homographs, that is, they are on the basis of word forms, not lemmas. Homographs, for instance, *mosimanyana* (small hole) and *mosimanyana* (small boy) or *mabele* (breasts) and *mabele* (sorghum) these are treated as the same item. Our calculations also do not make any distinctions on the basis of capitalisations. Therefore personal names such as *Masego* and *Thapelo* will not be distinguished from the common nouns *masego* (blessings) and *thapelo* (prayer). The calculations also include numbers such as the year 2006. Since the Setswana language does not use apostrophes, in our calculations we ignore apostrophes. This means that if there are some English words which use apostrophes in the corpus, the apostrophe will be taken as dividing two words. A similar approach is adopted in handling hyphens since there is no consistent manner of dealing with hyphenated words in Setswana orthography. We therefore treat hyphenated words as two distinct words.

To calculate keywords, we use WordSmith Tools' keyword program. The program

identifies "key" words in one or more texts. Keywords are words “whose frequency is unusually high in comparison with some norm” (Scott 2004-2006: 94). To calculate keywords frequency sorted wordlists are generated for a focus corpus (a corpus one is interested in) and for a reference corpus (a corpus that is larger than the focus corpus used as a reference/comparative corpus). The program conducts a statistical comparison between a wordlist of the focus corpus and that of a reference corpus to identify words which are key. The "key words" are calculated by comparing the frequency of each word in the wordlist of the focus corpus against the frequency of the same word in the reference corpus wordlist.

To compute the "key-ness" of an item, Scott (2004-2006: 97/8) points out that the program computes:

- its frequency in the small wordlist
- the number of running words in the small wordlist
- its frequency in the reference corpus
- the number of running words in the reference corpus

and these are cross-tabulated.

One way of explaining this process is to say:

1. Take two corpora or subcorpora: one large another small. The large one is a reference file, while the small one is the study corpus, the one we are interested in studying its lexical characteristics. A reference corpus has also been referred to as a “‘normative corpus’ since it provides a text norm (or general language standard) against which we can compare” (Rayson et al., 2004: 2).

2. Generate frequency lists from the two subcorpora.

3. Compare the frequency of each word in the study corpus against the frequency of a similar word in the reference corpus.

4. If a word is SIGNIFICANTLY MORE FREQUENT on the frequency list of the study corpus but SIGNIFICANTLY LOWER on the frequency list of the Reference corpus, list it as a possible definitive term (positive keywords).

5. If a word is SIGNIFICANTLY MORE FREQUENT on the frequency list of the study corpus and also SIGNIFICANTLY FREQUENT on the frequency

list of the Reference corpus, ignore it as uninformative/not defining the study corpus.

6. If a word is SIGNIFICANTLY LOWER on the frequency list of the study corpus and SIGNIFICANTLY MORE FREQUENT on the frequency list of the Reference corpus list it as a negative keyword.

7. If a word is SIGNIFICANTLY LOWER on the frequency list of the study corpus and it also SIGNIFICANTLY LOWER on the frequency list of the Reference corpus, ignore it as uninformative/not defining the study corpus.

The statistical tests include:

- the classic chi-square test of significance with Yates correction for a 2 X 2 table.
- Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against a reference corpus.

A word therefore identified as key if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist.
Culpeper (2002: 14) points out that keyness then "is a matter of being statistical unusual". Unusually *in*frequent key-words are called *negative key-words*. Unusually frequent key-words are called *positive key-words*. In this study we use the Log Likelihood test since it is considered better than the chi-square test of significance particularly when contrasting long texts or where one may have to deal with low counts of less than 5 log likelihood.

Log likelihood is calculated by constructing a contingency table as follows[20]:

Table 31: A contingency table

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

[20] http://ucrel.lancs.ac.uk/llwizard.html

The value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values). The values 'a' and 'b' are called the observed values (O), whereas we need to calculate the expected values (E) according to the following formula (also see Rayson et al, 2004):

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case N1 = c, and N2 = d. Therefore , E1 = c*(a+b) (c+d) and E2 = d*(a+b) (c+d). The calculation for the expected values takes account of the size of the two corpora, so we do not need to normalize the figures before applying the formula. We can then calculate the log-likelihood value according to the following formula:

$$-2\ln \lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating log-likelihood G2 as follows:

G2 = 2*((a*ln (aE1)) + (b*ln (bE2))). For the purposes of our experiments the likelihood measures are computed by the WordSmith software.

Kilgarriff notes that,

> *G*2 is a mathematically well-grounded and accurate measure of surprisingness, and early indications are that, at least for low and medium frequency words such as those in Daille's study, it corresponds reasonably well to human judgements of distinctiveness (Kilgarriff, 2001: 105).

The log likelihood statistic has been used before by Leech et al. (2001: 16) in the frequency analysis of the English language based on the British National Corpus. They chose the statistic for at least three reasons:

1. The statistic does not require the data to be distributed in a particular pattern

2. It does not over- or under-estimate the significance of a difference between samples unlike the Pearson chi-square test which has been shown to over estimate the importance of rare events.
3. It is insensitive to differences of size between two samples

The statistic has also been preferred by amongst others Rayson (2003) and Rayson et al. (2004).

The keywords extracted through keyword analysis characterise the domain of the text through their high occurrence in the study corpus compared to their frequency in the reference corpus.

The use of keywords for comparing corpora has been argued for by Sardinha (2000). It has been used by Culpeper (2002) for the analysis of words spoken by six characters in Romeo and Juliet. It has been preferred over Biber's multidimensional analysis (MDA) by Xiao and McEnery (2005) in genre analysis. They note that,

> MDA is undoubtedly a powerful tool in genre analysis. But associated with this power is complexity. The approach is very demanding both computationally and statistically in that it requires expertise not only in extracting a large number of linguistic features from corpora but also in undertaking sophisticated statistical analysis (Xiao and McEnery, 2005: 63).

They then demonstrate that using the keyword function of WordSmith Tools can achieve approximately the same effect as Biber's MDA. What attracted them to keyword analysis is that it is less demanding as WordSmith Tools can generate wordlists and extract keywords automatically.

Below we extensively use log likelihood measurement in keyword analysis to isolate words which characterise a genre or text type. Our aim is two fold: at one level we wish to measure whether a corpus compiled from various text types and genres is more attractive for lexicography if it could be found to generate unique words particular to each genre, which collectively capture the linguistic diversity present in every day language. A corpus that is designed in such a way as to capture the

linguistic diversity of a language would therefore be preferred over a corpus compiled from a single or a limited variety of genres. At another level, we hope that keyword analysis lists will by their distinctiveness communicate a related argument: that since the lists are dissimilar the course of lexicography cannot be served best by depending on a single text type for dictionary compilation, since a single text type will lead to the generation of a restricted lexicon. We use keyword analysis to extract genre specific lists, which Vintar (1999: 64) has argued that they "… can prove useful when studying the lexical specificity of a text or its terminological scope," an area we are currently investigating.

We first analyse the written part of the corpus and later look at the spoken components of the corpus. We measure the keywords of Science and technology, Politics, Poetry, Plays, Grammar books, Chatsite text, Religious text and the different parts of the Newspaper text. We provide the results below of only the top 100 most frequent tokens from the keyword lists derived from a variety of text types. We start with Science and technology keywords.

6.2.1 Keyword analysis of written components of the Setswana corpus

Table 32: Science and technology keywords

1. ict	23. diphelelo	45. dipuisano
2. botswana	24. wsis	46. mananeo
3. tshedimosetso	25. puso	47. selekanyo
4. hiv	26. megala	48. madirelo
5. ditirelo	27. pego	49. diteseletso
6. ditlhaeletsano	28. mafaratlhatlha	50. dirisa
7. aids	29. dikgaolo	51. ditshekatsheko
8. ditogamaano	30. tshwanetse	52. aforika
9. didirisiwa	31. badirisi	53. botlhokwa
10. karolo	32. ikemetseng	54. bobegadikgang
11. kitso	33. dikitsiso	55. kae
12. tiriso	34. lephata	56. boleng
13. metsi	35. dithuso	57. motlakase
14. bta	36. akaretsa	58. mafatshe
15. tlhabololo	37. molao	59. yunibesithi
16. sechaba	38. nang	60. dikgaolong
17. kgolagano	39. nyutlelia	61. bophara
18. mogare	40. lefatshe	62. dikhomputara
19. goromente	41. maphata	63. btc
20. inthanete	42. tsamaiso	64. workshop
21. maranyane	43. kgotsa	65. dirisiwa
22. botegeniki	44. tshekatsheko	66. seno

67. bokgoni	79. tlamo	91. letlhoko
68. borwa	80. ditlhwatlhwa	92. telecommunications
69. bolwetsi	81. rabies	93. batho
70. metswedi	82. dipatlisiso	94. thuto
71. tshono	83. technology	95. mafelo
72. lekalana	84. metlobo	96. patlisiso
73. seemo	85. badiri	97. tlhabolola
74. dikitso	86. taolo	98. kgaso
75. dipatisiso	87. itsholelo	99. dintlha
76. boitseanape	88. dirwa	100. icasa
77. megopolo	89. masome	
78. radio	90. mekgatlho	

The top 100 words in Table 32 characterise the science and technology text type which is a broad one. It includes medicine, computing, telecommunications and others. This variety of subfields is reflected in the variety of words from the different fields of science and technology captured in Table 32. We illustrate this variety by giving the words followed by their rank in brackets and if they are Setswana words we also offer English translation in brackets. Medical terms include *HIV* (4), *Aids* (7), *mogare* (18) (virus), *bolwetsi* (69) (disease), *rabies* (81); technology terminology includes *ICT* (1) (Information and Communications Technology), *BTA* (14) (Botswana Telecommunications Authority), *inthanete* (20) (internet), *maranyane* (21) (Science), *botegeniki* (22) (technology), *WSIS* (24) (World Summit on the Information Society), *nyutlelia* (39) (Nuclear), *motlakase* (57) (electricity), *dikhomputara* (62) (computers), *ditlhaeletsanyo* (6) (communications), *telecommunications* (92), *radio* (78), *ICASA* (100) (Independent Communications Authority of South Africa), *megala* (26) (telephones). Other words captured that are central to the area, albeit not in an obvious way include such words as *metsi* (13) (water), *tshekatsheko*, *ditshekatsheko* (44, 51) (investigations), *dipatisiso/dipatlisiso* (75/82) (research), and *didirisiwa* (9) (tools).

The Science and technology text type has comparatively higher levels of English words compared to the other lists that we will inspect later. Amongst these are *workshop* (64), *radio* (78), *rabies* (81), *technology* (83) and *telecommunications* (92). This is in part because some documents written in Setswana use English terms where the Setswana language does not have terms for certain science and technological concepts. In other instances the English words have been adopted into the Setswana language and spelt using Setswana orthography. Instances such as *inthanete* (20)

(internet), *nyutlelia* (nuclear) (39), *yunibesithi* (university) (59), and *dikhomputara* (62) (computers) are examples of Setswana words borrowed from English.

Next we conduct a similar experiment with politics text and the results follow in Table 33.

Table 33: Politics text keywords

1. bosetšhaba	35. karolwana	69. bothati
2. molaotheo	36. lefapha	70. dira
3. kgotsa	37. makoko	71. motlatsa
4. porofense	38. kgaso	72. kgololesego
5. peomolao	39. thulaganyo	73. melao
6. bommasepala	40. selegae	74. tiragatso
7. karolo	41. aforika	75. dithulaganyo
8. ditirelo	42. netefatsa	76. diphelelo
9. poresidente	43. mongwe	77. basarwa
10. molao	44. tonakgolo	78. ditšhelete
11. kokoano	45. tshedimosetso	79. ditheo
12. diporofense	46. dithata	80. fitlhelela
13. tshwanetse	47. tokololo	81. merero
14. mmasepala	48. baagi	82. tsweletsa
15. puso	49. komiti	83. mekgatlho
16. ditshwanelo	50. rephaboliki	84. dikgaolo
17. khansele	51. melawana	85. makgotla
18. tlhabololo	52. mametlelelo	86. maloko
19. khuduthamaga	53. bokgoni	87. kakaretso
20. molaotlhomo	54. setšhaba	88. lotseno
21. maikarabelo	55. akaretsa	89. tlamela
22. pusoselegae	56. kabinete	90. botlhe
23. lekoko	57. kgotlapeomolao	91. botho
24. ditlhopho	58. badiri	92. taolo
25. mmuso	59. anc	93. tshegetso
26. palamente	60. aretikele	94. ditshwetso
27. tshwanelo	61. borwa	95. maleba
28. mošwa	62. baemedi	96. mabakeng
29. khomišene	63. sepolotiki	97. tekatekano
30. tsamaiso	64. dikomiti	98. pegelo
31. ditokololo	65. sanetasi	99. tirisong
32. ditiro	66. demokerasi	100. palo
33. maikemisetso	67. botlhokwa	
34. kgotlatshekelo	68. botswana	

Politics deal with issues of governance (*puso* (15) or *mmuso* (25)), with the (*poresidente* (9)) president as the leader of cabinet (*kabinete* (56)), in (*palamente* (26)) parliament. The government runs through local governments (*bommasepala*

(6)), provinces (*diporofense* (12)), and council (*khansele* (17)). Government also deals with the enactment of laws. This is revealed by words such as *molaotheo* (2) (constitution), *peomolao* (5) (law enactment), *molao* (10) (law), *dithata* (46) (authority/powers), *melawana* (51) (statutes), *kgotlapeomolao* (57) (a gathering that creates laws), *aretikele* (60) (article) and *taolo* (92) (order). The broad area of politics also deals with all kinds of people, *setšhaba* (54) (a nation), *badiri* (58) (workers), *baemedi* (62) (representatives) and *Basarwa* (77) (the San/Bushmen) and ideals such as *ditshwanelo* (16) (rights), *demokerasi* (66) (democracy), *kgololesego* (72) (freedom), and *tekatekano* (97) (equality).

What also stands out from the top 100 politics words are Setswana terms which are used only in South African Setswana and not in Setswana used in Botswana. Below we give a comparative table that illustrates this phenomenon.

Table 34: South African Setswana politics terms and Botswana Setswana politics terms

SA Setswana word	Rank	Alternative Botswana Setswana	English
molaotheo	2	*molao-motheo*	constitution
bommasepala	6	-	local governments
poresidente	9	*tautona*	president
diporofense	12	*- (kgaolo)*	province
khomišene	29	*patlo-maikutlo*	commission
rephaboliki	50	*lefatshe*	republic
demokerasi	66	*puso ya batho ka batho*	democracy
ditšhelete	78	*madi*	money/funds

We will not explore the distinction between South African and Botswana Setswana any further here since it will be best to explore it across genres. However, we do raise it here since South African Setswana readers of this study may not pick this distinction while Botswana Setswana speakers may be surprised by the level of "South Africanisms" in the list. What must be remembered is that such a corpus output also reflects corpus input. It means that there are many texts from South African politics compared to Botswana politics. Most of the South African politics text is from the internet.

We now subject poetry text to keyword analysis and give the results of the top 100 words in Table 35.

Table 35: Poetry text keywords

1. ke	35. nka	69. tsala
2. gago	36. monna	70. gonne
3. wena	37. sona	71. mmoki
4. kgomo	38. maru	72. lorato
5. leboko	39. sekapuo	73. lala
6. yona	40. noka	74. tinkane
7. tau	41. ina	75. noga
8. kgosi	42. fatshe	76. tšhaba
9. pelo	43. maloba	77. tletse
10. motho	44. banna	78. kwena
11. tsona	45. mmopi	79. phologolo
12. pula	46. pitse	80. boroko
13. morwa	47. mariga	81. namane
14. aferika	48. bakgatla	82. kile
15. itse	49. gopola	83. duma
16. matlho	50. poko	84. tlhaga
17. naga	51. etsa	85. kgama
18. bosigo	52. ngwedi	86. pelong
19. ruri	53. tlou	87. sakeng
20. metsi	54. tlhe	88. moso
21. botshelo	55. meno	89. jewa
22. sala	56. tsatsi	90. boka
23. jaaka	57. bana	91. tlala
24. motse	58. maboko	92. mogatla
25. ngwana	59. lela	93. dithaba
26. morena	60. bogale	94. madiba
27. khunwana	61. nkwe	95. modimo
28. nna	62. tlhogo	96. kgarebe
29. lebokong	63. gareng	97. thobega
30. tlhaba	64. ditšhaba	98. rile
31. nonyane	65. phefo	99. mabana
32. mosadi	66. kgakala	100.kgwanyape
33. gaalelelwe	67. keledi	
34. nageng	68. mabele	

Setswana poetry is highly proverbial and rich with imagery. There was some concern that reducing it to a simple list will possibly completely obscure its sophistication and the images would be lost. This has not been the case. The images of animals, seasons and times, parts of the body, colours and other natural entities are revealed in the list. Wild animals used in Setswana poetry include amongst others *tau* (7) (lion), *tlou* (52) (elephant), *nkwe* (61) (tiger/leopard), *noga* (75) (snake), and *kwena* (78) (crocodile). Domestic animals include *kgomo* (3) (cow), *pitse* (46) (horse) and *namane* (81) (a calf). Times and seasons are captured by words such as *bosigo* (18) (night), *maloba*

(43) (some time ago) and *mariga* (47) (winter). Parts of the body used include *matlho* (16) (eyes), *meno* (55) (Teeth), *tlhogo* (62) (head), *mabele* (68) (breasts), and *pelong* (86) (in the heart). Other natural elements include *pula* (12) (rain), *naga/nageng* (17/34) (forest/wilderness), *maru* (38) (clouds), *ngwedi* (52) (moon), *phefo* (65) (wind), *tlhaga* (84) (grass), *dithaba* (93) (hills/mountains), and *madiba* (94) (lakes).

The poetry also deals with different persons. These include *kgosi* (8) (chief), *motho* (10) (person/individual), *ngwana/bana* (25/57) (child/children), *morena* (26) (lord/master), *monna/banna* (3644) (man/men), *Bakgatla* (48) (the tribe of the Bakgatla), *tšhaba/ditšhaba* (76/64) (nation/nations), *tsala* (69) (friend), *mmoki* (71) (poet) and *kgarebe* (96) (a young beautiful lady).

There is a detailed use of colour such as *khunwana* (27) (reddish brown colour in female animals) and *tlhaba* (30) (brownish colour in male animals).

One other common characteristic of Setswana poetry is the shortening of words by deleting their beginnings; prefix elision. This is reflected in two examples in the list: *ina* (41), a shortened version of *leina* (name) and *tsatsi* (56) a shortened form of *letsatsi* (day/sun) with the noun prefix *le-* in both cases elided.

There is also the reference to the unknown, divine or the imaginative creatures. These include *Mmopi* (45) (creator), *Modimo* (95) (God).

We have attempted to show that Setswana poetic language tends to use natural images such as wild and domestic animals, seasons and times and natural elements.

We now turn to another type of creative work, plays. We subject plays to keyword analysis. The results follow in Table 36.

Table 36: Plays text keywords

1. gae	7. gago	13. rra
2. tlaa	8. rakgomo	14. mmamoilwa
3. thotseditlotse	9. mmaselepe	15. rothodilapule
4. mma	10. borutuse	16. rapeipi
5. modiri	11. wena	17. ntesang
6. rona	12. matshediso	18. khumo

19. tibe	47. butiki	75. sibinjolo
20. kgonamanaba	48. jojina	76. oteng
21. motimedi	49. radipodi	77. mmabatho
22. kegakilwe	50. khutsafalo	78. thotobolo
23. bojosi	51. mmadikatse	79. nteseng
24. ngaka	52. kampo	80. photo
25. kesara	53. wa	81. mmabogobe
26. ntlale	54. bua	82. jaana
27. mofalotsi	55. ela	83. motlhalefi
28. kana	56. joshi	84. rrabogobe
29. kasiuse	57. setsumpa	85. nombini
30. eng	58. tawane	86. mmalefa
31. lona	59. fokolengwe	87. lethosa
32. nna	60. mmamitlwe	88. seemo
33. mmelegi	61. motshwarateu	89. antoniuse
34. pulane	62. zuu	90. simane
35. jaanong	63. ruri	91. mmanyai
36. mosenyi	64. maswe	92. mokgalo
37. seikanyeng	65. mogologolo	93. tshudube
38. amantle	66. ipuseng	94. letsoro
39. tlhoriso	67. ngwanaka	95. nka
40. boikobo	68. batla	96. lerato
41. makgoropetsa	69. morobi	97. megare
42. rapuo	70. sekei	98. montsana
43. tefo	71. mmadipodi	99. lebutle
44. mpho	72. kedibone	100. moutlwatsi
45. mogapinyana	73. itse	
46. ditshele	74. kgotso	

Plays are the dramatisations of people's stories. In writing, these are rendered with personal names followed by an individual's written words. This results inevitably in a high repetition of speakers' names in the whole text. The top 100 words list in Table 36 therefore has a large number of personal names, 82 in all. These are highlighted in the above list. Only eighteen words are not personal names. These include *gae* (1) (home), *tlaa* (2) (will), *mma* (4) (mother (of)), *rona* (6) (us), *gago* (7) (yours), *wena* (11) (you). Apart from observing that the overwhelming majority of the top 100 Plays' keywords are personal names, we cannot adequately characterise the Plays' keywords. To characterise the Plays' keywords adequately, strategy of dealing with personal names in such a way that they do not interfere with the counts was deleted. The speaker's names as metatext were treated as metatext and marked up in such a way as to exclude them from the counts. The following example illustrates the mark-up strategy adopted.

<c>Bothata</c> A re o lomiwa ke eng?

<c>Thekiso</c> Ke ka bo ke akga loleme fa ke ka go raya ka re ke itse se se mo jang.

<c>Bothata</c> Tlhokomologa tseo ngwanaka a re robale. Gongwe o itse se a se lwelang, o tlaa itlhalosa fa a na le kgang. Tshu! ke šele jang. Letsatsi le sala le tlhola le kgwisa kolobe diphulo. Tima lebone foo mma.

After the play's text was marked-up the frequency counts were run and the results of the experiment follows in Table 37.

Table 37: Plays text keywords with names treated as metatext

1.	ke	36,342.36	39.	batla	1,327.55
2.	re	7,602.77	40.	mosadi	1,292.57
3.	lo	6,885.52	41.	nka	1,208.66
4.	ka	6,726.40	42.	jaanong	1,177.17
5.	ga	6,704.10	43.	gore	1,058.94
6.	me	6,649.80	44.	kgosi	1,053.96
7.	go	5,076.64	45.	ruri	1,030.03
8.	se	4,916.64	46.	mosimane	1,021.81
9.	tla	4,728.61	47.	a	994.54
10.	wena	4,063.59	48.	lerato	987.67
11.	gago	4,048.78	49.	tsala	981.15
12.	bona	3,416.03	50.	sa	971.52
13.	itse	2,911.68	51.	bo	956.12
14.	le	2,900.26	52.	na	948.85
15.	rra	2,841.73	53.	iwa	939.71
16.	eng	2,738.55	54.	rona	931.70
17.	mma	2,640.35	55.	tsamaya	888.35
18.	nna	2,639.68	56.	jaana	880.69
19.	monna	2,611.53	57.	kae	878.94
20.	yona	2,347.86	58.	kana	878.54
21.	mo	2,164.68	59.	tle	860.30
22.	ha	2,127.68	60.	raya	859.16
23.	gone	2,070.49	61.	gagwe	849.10
24.	wa	2,054.86	62.	ne	848.03
25.	fa	1,801.20	63.	ona	785.32
26.	yo	1,781.09	64.	nnyaa	780.27
27.	motho	1,650.96	65.	ngaka	747.39
28.	bua	1,603.06	66.	fela	724.38
29.	ngwana	1,583.32	67.	botshelo	717.88
30.	pelo	1,570.38	68.	neng	714.21
31.	tsena	1,528.03	69.	ye	710.11
32.	ntse	1,447.01	70.	sepe	696.98
33.	lona	1,445.25	71.	matlho	695.55
34.	utlwa	1,429.64	72.	rata	670.24
35.	sona	1,396.37	73.	siame	645.96
36.	ngwanaka	1,379.22	74.	koko	642.76
37.	tsona	1,371.96	75.	kete	632.11
38.	tlaa	1,332.70	76.	tlhe	619.46

77.	kampo	618.20
78.	sengwe	604.83
79.	be	602.51
80.	ee	590.91
81.	tshega	586.26
82.	fano	565.70
83.	sentle	563.01
84.	tswa	557.95
85.	tlile	548.69
86.	twe	548.02
87.	ao	545.37
88.	ene	540.97
89.	utlwile	534.76

90.	botlhoko	528.49
91.	mosetsana	522.90
92.	tlhogo	522.21
93.	jang	519.62
94.	bosigo	514.07
95.	setse	501.38
96.	ise	491.63
97.	ijoo	486.10
98.	ena	485.28
99.	dikgomo	483.85
100	ntlong	475.17

Since speakers have to refer to themselves and each other and not always through personal names, pronouns are common. Speakers also refer to the space within which events take place. The results show that most of the top 100 keywords are functional words amongst these being a variety of concords, auxiliary verbs used in negative constructions, pronouns, demonstratives such as *ke* (1) (I), *re* (2) (we), *lo* (3) (you (plural)), *me* (6) (mine), *go* (7) (there (existential)), *se* (8) (it, this), *wena* (10) (you (singular)), *gago* (11) (yours), *mo* (21) (this), *ha* (22) (here, give), *fa* (25) (here, give), *yo* (26) (this one), *tsona* (37) (them), *a* (47) (of, he, she), *bo* (51) (it), *gagwe* (61) (his (possessive)), *ona* (63) (it), *sengwe* (78) (something), *fano* (82) (here), *ene* (88) (him, her), *ke* (1) (I, he, she, it), and *ena* (98) (him, her). These results are consistent with the findings of Allwood (1998) who found out that pronouns made up over 25% of the Swedish spoken corpus.

Other terms found amongst the top 100 words are prepositions, conjunctions, possessive concords such as *ka* (4) (with), *ga* (5) (of), *wa* (24) (of), and *sa* (50) (of). We also have interjections such as *nnyaa* (65) (no), *ee* (80) (yes), *ao* (87) (wow!), and *ijoo* (97) (a cry for help or a cry of surprise). Other terms include conjunctions such as *le* (14) (and), *gore* (43) (that, where), *kana* (58) (or), *kampo* (77) (perhaps, or). There are auxiliary verbs as well, such as *tlaa* (38) (will), *nka* (41) (can, may), and *neng* (68) (was).

The list also includes adverbs such as *kae* (57) (where), *fela* (66) (only), *be* (79) (then), and *jang* (93) (how) a variety of verbs such as *tla* (9) (come), *ha* (22) (here, give), *raya* (60) (say, tell, mean), *ne* (62) (was), and *twe* (86) (said).

Since plays are about human relations, their conflicts and how they relate to each other, what also stands out is kinship terms and other words which people in dialogues use to address each other, for instance, *rra* (15) (father of/sir), *mma* (17) (mother of/madam) *monna* (19) (man/husband), *motho* (27) (person), *ngwana* (29) (child/baby), *ngwanaka* (36) (my child), *mosadi* (40) (woman/wife), *kgosi* (44) (chief), *mosimane* (46) (boy), *tsala* (49) (friend), *ngaka* (65) (traditional doctor), *koko* (74) (granny) and *mosetsana* (91) (girl).

Other nouns include *pelo* (30) (heart), *botshelo* (67) (life), *matlho* (71) (eyes), *tlhogo* (92) (head), *bosigo* (94) (night) and *dikgomo* (99) (cattle).

Koko (74) in Setswana is an ambiguous word since it could mean "chicken" or it could be "a verbal knock" at the door or a short form for "*nkoko*" (grandmother). A concordance analysis of the whole corpus revealed 249 concordance lines. Two instances of these are with *koko* as a knock at the door. This is illustrated in the following example:

	Concordance
14	tlhola mo go e. (Ba ikaba ka matlo). : Koti! Koko! A go mongwe mo ntlong. Nte ke leke
15	'a bone. 0 emisa fa sellhareng sa morula) Koko! A go na le batho. Mme dinaonyana

Fifteen instances of *koko* were of the meaning of "chicken". We sample a few of these below:

	Concordance
77	le ka namana. O lese go tlabatlaba jaaka koko e batla go baya lee. Rrago ke yo le
78	. . .! : Ga re iketle , tota fa o tlabatlaba jaaka koko e batla go baya lee jaana wa re go
79	O fitlhela ba tlola ba tl~h~t.laha inaka koko e bona mmidi mo lebot- Tota ke eng ka
80	me lo ganelela kwa teng. Ke korakora jaaka koko e fesiwa ke lee. Mong wa me, ke
81	teng lo tla fitlhela go builwe. : Re tla etsa koko e gopotse mae; Re tla fota re obile
82	ka tladi, le go ba foufatsa ba nna jaaka koko e jwetswe ke noga, le go ba isa le naga,
83	a bonala gore mo bokgarebeng jwa gagwe, koko e ne e le senatla. : Mo tseleng e e

The overwhelming uses of "*koko*" are as a shortened version of *nkoko* (granny). We

offer a few concordance lines to illustrate such usage:

	Concordance
234	go tlhapise. : Ee, ngwana wa tsala ya me. : Koko, tsenya seatla pele mo metsing 0 utlwe
235	Mafoko a ga moruti a file koko tsholo- felo. O supeditswe gore go na
236	: Keresemose ya monongwaga e tsiseditse koko tsholofelo. ngwa- ga ono kewa poloko
237	a a rapeletsweng. MMASEI.EPE: Moruti koko Tshotseditlotse o bua boammaaruri fa
238	ya Modimo. Fela ga ke ye gope! Nnyaya koko, utlwa kopo ya me! E setse e nna
239	ntlogele ke itlotlele le koko wena. Jaanong koko wa bona lebole le lame le, le maatla mo
240	gago kgotsa dikoko- mane tsa gago gore koko wa bonangnang o di tsenye ? Tlhalosa
241	mm' mme le bongwaanake ba mo tsaa jaaka koko wa bone. Ba a mo tlotla theta. : Lo
242	fa nka ikgaoganya le kereke ya Modimo. : Koko, wa ikgaoganya le poloko ya mowa wa
243	a ile go nna kwa ga gabr Molefi, rona koko wa rona e tla nna mang? : Wena le
244	tlhe bathong. : A ke 0 ntlogele ke itlotlele le koko wena. Jaanong koko wa bona lebole le

What we have shown so far is that while Plays text deals with a variety of thematic issues, it does so through inter-personal relations revealed through the use of names, pronouns and interrogative indicators.

Having looked at Plays texts, focus is shifted to grammar texts. Grammar texts are of great interest since learners at different levels study Setswana grammar. Grammatical texts are also relevant in that they constitute technical writing in that they deal with specialised linguistic terminology. We give the results of keyword analysis of grammar texts in Table 38.

Table 38: Grammar texts keywords

1. latelang	18. ditumanosi	35. mefuta
2. dikao	19. godimo	36. ditlhaka
3. sekao	20. buisa	37. matlhaodi
4. puo	21. dikgomo	38. dikwalwa
5. kwala	22. madiri	39. mokgwa
6. lediri	23. mosimane	40. matlhalosi
7. kgotsa	24. temana	41. sengwe
8. naya	25. kgomo	42. tiriso
9. mafoko	26. bokao	43. lereo
10. maina	27. ithuta	44. arabe
11. kopulatifi	28. dipolelong	45. felo
12. dipolelo	29. polelo	46. mosetsana
13. dipotso	30. leboko	47. dikaong
14. popego	31. mmoki	48. dingwe
15. farologaneng	32. baithuti	49. dikutu
16. dirisa	33. monna	50. poko
17. mowa	34. barutwana	51. moela

52. tumiso	69. thutapuo	86. loleme
53. tlhaloso	70. bonolo	87. sefonetiki
54. morutabana	71. kutu	88. lemoga
55. tekolo	72. tlotlofoko	89. bala
56. tlhagisa	73. modumo	90. ngwana
57. medumopuo	74. ntlha	91. medumo
58. ditumammogo	75. letiro	92. temaneng
59. metsi	76. ditaelo	93. kganetsong
60. dikelo	77. ditlamorago	94. tumarinini
61. dipopi	78. sekameng	95. phologolo
62. tlhogo	79. mathusamadiri	96. dikwalo
63. sediri	80. setlhare	97. dipounama
64. tumisong	81. mosadi	98. jalojalo
65. matlaleletsi	82. maleba	99. nngwe
66. tlhaola	83. motho	100.polelong
67. leina	84. popi	
68. segalo	85. dirisiwa	

The top 100 keywords of grammar texts are dominated by linguistic terms. The area of linguistics is a specialised one, particularly in the Setswana language, with grammar terms being highly specialised to the genre and rarely occurring in other genres. Through keyword analysis we have extracted the following terms: *madiri/lediri* (22/6) (verbs/verb), *kopulatifi* (11) (copulative), *ditumanosi* (18) (vowels), *ditlhaka* (36) (letters/alphabet), *matlhaodi* (37) (adjectives), *matlhalosi* (40) (adverbs), *lereo/leina* (43/67) (name), *medumopuo* (57) (speech sounds), *ditumammogo* (58) (consonants), *tlhogo* (62) (prefix/head/subject), *sediri* (63) (subject), *matlaleletsi* (65) (objects), *thutapuo* (69) (a grammar book), *segalo* (68) (diacritic), *kutu* (71) (stem), *modumo* (72) (sound), *mathusamadiri* (79) (auxiliary verbs), *sefonetiki* (87) (phonetics), *tumarinini* (94) (palatalisation), *dipounama* (97) (lips/bilabial) and many others.

The Grammar texts include numerous exercises with instructions for students. Such instructions are reflected in the list words such as *latelang* (1) (following), *sekao* (3) (an example), *kwala* (5) (write), *naya* (8) (give), *buisa* (20) (read), *arabe* (44) (answer), *tlhaola* (66) (separate), *ditaelo* (76) (instructions), *lemoga* (88) (identify, realise), and *bala* (89) (read).

We now look at Arts and culture texts. These texts are from the Setswana newspaper, *Mokgosi*. They are about music, art, and a variety of cultural events.

Table 39: Arts & culture text keywords

1. mmino	35. diseko	69. ditlhako
2. pina	36. kgaisano	70. bontle
3. dipina	37. folaga	71. ipolelela
4. alebamo	38. motshwantshi	72. jaanong
5. gagwe	39. ngwao	73. mmala
6. opela	40. sekoleng	74. bua
7. baopedi	41. meropa	75. ditsala
8. senyatso	42. mochankana	76. mogakolodi
9. batshwantshi	43. bareki	77. disco
10. moopedi	44. mafohle	78. bataki
11. monate	45. logong	79. jazz
12. puna	46. botaki	80. papetlana
13. bajibareki	47. rata	81. dilo
14. ditshwantsho	48. kopelo	82. black
15. ditshupo	49. dira	83. opelwa
16. thapong	50. machesa	84. tshameka
17. diletso	51. maitisong	85. baboki
18. moopelo	52. tshwantsha	86. game
19. lumumba	53. steers	87. lokwalo
20. modimakwane	54. olebogeng	88. tonki
21. eric	55. sespo	89. utlwa
22. lorato	56. motho	90. jese
23. poko	57. matheke	91. banjo
24. barati	58. setiko	92. tsamaya
25. baji	59. joyce	93. wame
26. botshelo	60. mbaki	94. maitiso
27. mosadi	61. ipotsa	95. monna
28. bile	62. tota	96. tjiyapo
29. campbell	63. baletsi	97. modimo
30. setlhopha	64. talente	98. lelwapa
31. banna	65. basadi	99. kwaito
32. vivian	66. bogole	100. dinkgwana
33. ngwana	67. thata	
34. setso	68. boikanyo	

The Arts and Culture genre include texts primarily from music, art (drawing and painting), and other artistic expressions. From the area of music we find different types of music: Disco (77) *Jese* (90) or Jazz (79) and *Kwaito* (99). There are also names of musicians and bands such as *Senyatso* (8), *Puna* (12), Eric (21), *Botshelo* (26), *Machesa* (50), *Sespo* (55), *Matheke* (57), and *Banjo*[21] (91). There are also music related nouns and verbs such as *mmino* (1) (music), *pina/dipina* (2/3) (song/songs),

[21] Banjo as used here does not refer to the name a musical instrument, but a Botswana jazz musician known as Banjo Mosele.

alebamo (4) (album), *opela/opelwa* (6/83) (sing/sung), *baopedi/moopedi* (7/10) (musicians/musician), *diletso* (17) (musical instruments), *moopelo* (18) (Music), *setlhopha* (30) (band/group), *meropa* (41) (drums, also the name of a Jazz club in Gaborone), *baletsi* (63) (players) and *Maitiso* (94) (evening intertainment/also the an performance arts hall in Gaborone). The arts are revealed by the words *ditshwantsho* (14) (pictures/photos), *ditshupo* (15) (exhibitions), *Thapong* (16) (the name of an association of artists), *motshwantshi* (38) (an artist/one who draws), *botaki* (46) (art), *tshwantsha* (52) (draw), *talente* (64) (talent) and *mmala* (73) (colour). Other cultural terms are *poko* (23) (poetry), *setso* (34) (custom), *ngwao* (39) (culture), *baboki* (85) (poets), and *dinkgwana* (100) (clay pots).

Keyword analysis here reveals a variety of artistic and cultural terminology from the Arts and culture text.

We now turn to a different kind of text, chat-site text downloaded from the internet. Chat-site text is interesting since it is "raw" and "dirty" text; raw in having not been subjected to any editorial policy, especially when compared to grammar text and plays text that have already been analysed and it is dirty in that it includes misspellings, English words and colloquialisms. The results of keyword analysis on chat-site text follow in Table 40.

Table 40: Chat-site text keywords

1. the	18. not	35. my
2. to	19. this	36. people
3. posted	20. have	37. will
4. i	21. we	38. who
5. you	22. they	39. like
6. and	23. all	40. message
7. on	24. with	41. just
8. of	25. but	42. know
9. at	26. what	43. here
10. by	27. edumela	44. there
11. that	28. your	45. about
12. is	29. if	46. from
13. 2002	30. do	47. or
14. in	31. so	48. email
15. are	32. as	49. think
16. it	33. topic	50. com
17. for	34. can	51. oct

52. when	69. how	86. only
53. be	70. would	87. because
54. no	71. board	88. those
55. our	72. back	89. want
56. am	73. nov	90. should
57. botswana	74. he	91. page
58. was	75. chat	92. love
59. get	76. has	93. time
60. home	77. guys	94. batswana
61. us	78. then	95. good
62. out	79. click	96. way
63. their	80. some	97. make
64. them	81. 2003	98. now
65. up	82. aids	99. dont
66. its	83. man	100.even
67. say	84. an	
68. why	85. other	

What sets the Chat-site language apart is its broad use of English words and obvious internet terminology. In all of the top 100 keywords none of the words are in Setswana. The internet terminology include amongst others, *posted* (3), *message board* (40, 71), *email* (41), *com* (50), *home page* (60, 91), *chat* (75), *click* (79) and the name of the chat-site, *Edumela* (27).

There are other words which are common in dialogues. Such words include pronouns such as *I* (4), *you* (5), *that* (11), *this* (19), *we* (21), *they* (22), *my* (35), *there* (44), *our* (55), *their* (63), *them* (64), *its* (66), *he* (74), and *those* (88). These are similar to those commonest English words in informal English speech showing that chat-site language has high instances of English and it is characterised by informality (Leech et al., 2001).

There are also interrogatives such as *what* (26), *who* (38), *when* (52), *why* (68) and *how* (69).

In terms of the subjects that are handled in this chat-site it appears that the top 100 words reveal very little save for words such as *people* (36) *Botswana* (57), *guys* (77) *Aids* (82), *man* (83) *love* (92), *Batswana* (94) which hint at the discussion on Botswana and Batswana, relationships and diseases such as Aids.

What is clear therefore from Chat-site text is that it is a text with high levels of

English words. Like spoken language and Plays text it uses many pronouns and to better characterise the kind of subjects handled in the chat-site, one would have to analyse more than the top 100 keywords.

We now look at the newspaper news section of the corpus. The news are largely from the *Mokgosi* newspaper and *Naledi*, the *Mmegi* newspaper insert. Like the previous text types we subject it to keyword analysis and we give the results in Table 41.

Table 41: News text keywords

1. botswana	35. tsedi	69. ikemetseng
2. e	36. lephata	70. seemo
3. aforika	37. bobegadikgang	71. bnf
4. puso	38. domi	72. tebelopele
5. mokgosi	39. tsamaiso	73. khama
6. batswana	40. kompone	74. lekalana
7. phathi	41. leno	75. didirisiwa
8. mafatshe	42. gaborone	76. phuthego
9. banana	43. batlhophi	77. letlhoko
10. aids	44. mopalamente	78. maloko
11. tlhalositse	45. itsholelo	79. ditshwanelo
12. ditlhopho	46. makgotla	80. boletse
13. lekgotla	47. mogae	81. diteseletso
14. bomme	48. komiti	82. francistown
15. hiv	49. santse	83. dikompone
16. tautona	50. bosheng	84. bagwebi
17. bta	51. tona	85. ndlovu
18. mogare	52. ditlhabololo	86. gatwe
19. ditlhaeletsano	53. btc	87. kgaolong
20. setshaba	54. zimbabwe	88. mafatsheng
21. babereki	55. setlhopha	89. sechaba
22. ditirelo	56. bone	90. mmino
23. masome	57. lenaneo	91. bush
24. goromente	58. kgaisanyo	92. molefhabangwe
25. lefatshe	59. kare	93. basarwa
26. ngwaga	60. dithuso	94. mafatshefatshe
27. domkrag	61. akaretsa	95. dipuisanyo
28. borwa	62. sepolotiki	96. mmaraka
29. ict	63. pego	97. dikgaolo
30. amerika	64. batshameki	98. 2016
31. seka	65. ditogamaano	99. babegadikgang
32. diphathi	66. ebile	100. ipapatso
33. applications	67. palamente	
34. gotwe	68. mmegi	

Newspaper text keywords cover a broad spectrum of subjects just as news text does.

There are political terms including political parties such as *Domi* (38) or *Domkrag*[22] (27), BNF[23] (71), *ditlhopho* (12) (elections), *batlhophi* (43) (voters), *mopalamente* (44) (Parliamentarian) *tona* (51) (minister/big/large), *sepolotiki* (62) (politics) *palamente* (67) (parliament), and *ipapatso* (100) (campaign), and others. There are also political personalities such as (President) *Mogae* (47) and (President) *Bush* (91), (Minister) *Molefhabangwe* (92) and the (Vice President of Botswana, Ian) *Khama* (73), all who have been newsmakers in Botswana. Botswana has also been promoting its national vision 2016, *Tebelopele 2016* (72/98) which articles how the nation desires to be by the year 2016. Other terms are clearly from the business sector. These include amongst others *itsholelo* (45) (economy), *diteseletso* (81) (licenses), *dikompone* (83) (companies) and *mmaraka* (96) (market). Other terms are technological. These include *ditlhaeletsano* (19) (communications), BTA[24] (17), ICT[25] (29).

News keywords therefore cover a diversity of subjects, just as newspapers themselves cover a variety of subjects.

We now look at Religious text which comprises mainly of Christian text. The top 100 keywords follow from this text type below in Table 42.

[22] Nicknames for the Botswana Democratic Party

[23] Botswana National Front

[24] Botswana Telecommunications Authority

[25] Information Communication Technology

Table 42: Religious text keywords

1. morena	35. gone	69. molemo
2. modimo	36. rona	70. moreneng
3. gago	37. motlhanka	71. ditirafalo
4. dafita	38. kae	72. diatleng
5. baiseraele	39. bafelesita	73. baroma
6. iseraele	40. selefera	74. baikepi
7. morwa	41. luke	75. bopelotlhomogi
8. jerusalema	42. morafe	76. morwawe
9. lefatsheng	43. farao	77. direla
10. juta	44. balefi	78. aborahame
11. jesu	45. bone	79. bosula
12. kgosi	46. yotlhe	80. mathaio
13. bomorwa	47. medimo	81. moweine
14. botlhe	48. felong	82. bong
15. egepeto	49. ditlhabelo	83. moporofeti
16. boitshepo	50. bajuta	84. boipontsho
17. saule	51. bophelo	85. legodimo
18. keresete	52. baebele	86. losika
19. lefatshe	53. phiso	87. tshiamo
20. moperesiti	54. legodimong	88. hesekia
21. tsotlhe	55. aletare	89. letlole
22. jakobe	56. jeso	90. sabata
23. setlhabelo	57. bomorwawe	91. gagwe
24. arone	58. masomosomo	92. tatlhego
25. gouta	59. samuele	93. tente
26. batlhanka	60. johane	94. jehofa
27. baperesiti	61. bosakhutleng	95. otlhe
28. boleo	62. raya	96. lotlhe
29. salomo	63. nne	97. kajeno
30. babele	64. tšhaka	98. joabe
31. dikgosi	65. kgolagano	99. jesaya
32. dinyaga	66. diane	100.lotso
33. jobe	67. tshupelo	
34. jeremia	68. khutleng	

The religious keywords analysis reveals the dominance of Christian text in our subcorpus. The top 100 keywords include books of the Bible such as *Jakobe* (22) (James), *Dikgosi* (31) (Kings), *Jobe* (33) (Job), Jeremia (34), Luke (41), *Samuele* (59) (Samuel), *Johane* (60) (John), *Diane* (66) (Proverbs), *Ditirafalo* (71) (Chronicles), *Baroma* (73) (Romans), *Mathaio* (80) (Matthew), and *Jesaya* (99) (Isaiah). There are also Biblical figures such as *Dafita* (4) (David), *Baiseraele* (5) (Isralites), *Jesu/Jeso* (11, 56) (Jesus), *Saule* (17) (Saul), *Keresete* (18) (Christ), *Moperesiti* (20) (priest), *Arone* (24) (Aaron), *Bafelesita* (39) (Philistines), *Bajuta* (50) (Jews), *Aborahame* (78) (Abraham) and *Moporofeti* (83) (a prophet). Other religious terms include *Morena* (1)

(Lord), *Modimo* (2) (God), *Medimo* (47) (gods), *morwa* (7) (Son (of God)), *kgosi* (12) (King/chief), *boitshepo* (16) (holiness), *setlhabelo/ditlhabelo* (23/49) (sacrifice/sacrifices), *boleo* (28) (sin), *selefera* (40) (silver), *legodimo/legodimong* (85/54) (heaven/heavenly), *aletare* (55) (altar), *baikepi* (74) (sinners), *bosula* (79) (wickedness), *tatlhego* (92) (lostness) and *Sabata* (90) (Sabbath).

The religious terminology is unique to the genre of religion, and in this case particular to the Christian religion.

Having looked at the words that characterise the area of religion we now look at corpus components of spoken text and isolate their keywords.

6.2.2 Keyword analysis of spoken components of the Setswana corpus

This part of the corpus comprises transcribed live football commentaries from Radio Botswana and sport report on a variety of games. We will analyse, Call-in, face-to-face dialogue, classroom interactions, Hansard, radio interviews, open radio programming, Religion and Sport text.

We begin our analysis with Call-in text. The data used in this analysis is from three call-in programs *Moremogolo*, *Maokaneng, Phutha-ditšhaba* and *A re bueng*. The topic on *Moremogolo* was on the offensiveness of cellphone use and how they lead to deceptiveness since speakers claim to be in certain locations when they in fact are far from them. *Phutha-ditšhaba* dealt with elections, precisely citizens' readiness to vote. *A re bueng* dealt with how certain children terrorise parents by making unreasonable demands and when their demands were not honoured they threaten suicide. *Maokaneng* deals with the role of the media in elections. We mention the subjects handled in these programs to shed light on the words in table 43.

Table 43: Call-in text keywords

N	Keyword	Keyness
1	ee	5598.792969
2	ke	4650.180664
3	re	3571.413818
4	gore	2647.37085
5	rra	2498.179688
6	ko	2351.857422
7	an	2099.151855

8	hello	1742.438477
9	jaanong	1212.748047
10	nnyaa	1009.634583
11	tankie	720.6206055
12	raya	703.6761475
13	mma	637.4937744
14	ntse	590.5014648
15	ehe	589.7432861
16	lapeng	564.6315918
17	okay	530.8041382
18	ngwana	524.657959
19	cellphone	523.3270874
20	ipolaya	512.5631714
21	rre	507.0570068
22	kana	499.7451172
23	fa	497.8006287
24	bo	496.7189941
25	lebogile	470.307373
26	nna	461.8958435
27	radio	452.9598694
28	tla	452.5892639
29	bye	428.1827087
30	le	415.0164795
31	bua	403.5699768
32	teng	403.0974426
33	wena	392.680542
34	bona	391.37854
35	bana	389.4447021
36	jang	387.3610535
37	tota	367.9024658
38	mme	366.5827637
39	fela	354.6725464
40	batsadi	344.4155884
41	ntate	340.0013428
42	leng	335.1376953
43	rona	333.7392578
44	hela	325.0430603
45	gone	310.8773499
46	eng	305.8117371
47	nkgonne	304.1421204
48	itse	295.0811462
49	mo	278.6512756
50	utlwa	270.1494751
51	go	269.0933533
52	ga	262.1260986
53	batla	249.0783386
54	na	248.7524261
55	mogaetsho	248.1150818
56	jaana	245.1821289
57	pule	243.3664246
58	gongwe	238.5174866
59	ba	229.0778656
60	rraagwe	217.4441681
61	leboga	212.4922943
62	ngwanaka	210.8857269
63	matlhophelong	206.3597717
64	kae	196.0868378
65	moso	195.6818695
66	ka	194.9997406
67	ra	190.0220947
68	ha	188.3122253
69	chris	186.6708221
70	gago	183.3898163
71	golo	182.3194733
72	reng	180.725174
73	phutha	173.7875977
74	se	167.048996
75	lo	166.1824951
76	tsaya	164.6233063
77	moremogolo	163.4777679
78	ye	158.3822937
79	ditshaba	156.4868469
80	dilo	152.937088
81	ehee	148.8966675
82	tsamaya	148.3753815
83	semang	143.8077393
84	kgang	142.1442719
85	gaetsalwe	142.0330048
86	gompieno	140.5166931
87	sengwe	140.3969421
88	utlwile	136.7749939
89	rraetsho	135.7851563
90	nale	132.7247314
91	rekela	132.4711914
92	motsadi	127.2045517
93	thupa	124.4304581
94	kgona	121.6762238
95	dithato	121.6504059
96	tle	121.1425247
97	bue	120.5993958
98	setlhogo	120.2457581
99	mozeregwa	119.606102
100	campaign	117.7104187

The list gives evidence of words common in dialogue by the use of such words as *ee* (1) (yes), *nnyaa* (10) (no) *hello* (8), *tankie* (11), or *ke lebogile* (2/25) (thank you) and

bye (29).

Pronouns in direct communications rank amongst the most frequent. These are *nna* (26) (me), *wena* (33) (you), and *rona* (43) (us).

A variety of words which mark interrogatives appear in the list indicating that the presenter asks a series of questions to the callers. These include words *jang* (36) (how) *leng* (42) (when), *eng* (46) (what), *kae* (64) (where/how much).

Words that reflect respectful dialogues between individuals *nkgonne* (47) (elder brother/sister), *mogaetsho* (55) (colleague), *batsadi* (40) (parents) *rraetsho* (89) (sir), *motsadi* (92) (parent), appear amongst the top 100.

Other words hint at the topics discussed in the dialogues. Some of these words are *cellphone* (19), *ipolaya* (20) (commit suicide), *matlhophelong* (63) (voting stations), *rekela* (91) (buy for) and *thupa* (93) (lash/stick).

Included in the spoken subcorpus are the face-to-face dialogues. These are recordings of family interactions. The text from these is small. Consequently only 71 keywords have been extracted instead of 100.

Table 44: Face to face dialogue keywords

N	Keyword	Keyness
1	re	490.53
2	ee	338.58
3	nnyaa	208.44
4	ke	203.4
5	ko	142.62
6	ba	137.33
7	lo	129.81
8	kgosing	117.72
9	raya	106.36
10	ehe	99.175
11	plaka	91.315
12	fela	80.89
13	mmathapelo	78.269
14	tswaletswe	71.928
15	dikgomo	70.978
16	mm	65.223
17	tlhotse	64.623
18	hela	64.497
19	tswetswe	58.268
20	malutu	57.128
21	kana	55.318
22	ga	54.396
23	win	52.177
24	ijoo	51.33
25	gana	51.03
26	bogadi	48.315
27	lona	48.075
28	talela	47.176
29	apeetse	47.176
30	fologa	44.879
31	welang	43.33
32	tshekong	42.774
33	montshonyana	42.626
34	tswale	41.099
35	kere	39.627
36	mpoleleleng	39.132
37	leteitsi	39.132

38	theogela	39.014
39	molaodi	38.759
40	jaana	36.09
41	tsee	35.727
42	sekgoa	34.125
43	gatwe	32.629
44	ha	31.573
45	maabane	31.263
46	dikologa	30.98
47	golo	30.411
48	itshwere	29.917
49	mmm	29.583
50	bo	29.582
51	eng	29.426
52	mpolelele	28.562
53	twe	28.242
54	fa	28.138
55	jaanong	28.088

56	tv	27.887
57	tleng	27.887
58	kwano	27.771
59	tilwe	27.693
60	gore	27.684
61	nyalwa	27.392
62	gogwa	26.926
63	tweng	26.891
64	batla	26.88
65	nna	26.86
66	phakela	25.121
67	nyetse	25.121
68	bogosi	25.118
69	phoso	24.326
70	yo	24.23
71	letse	24.115

As in many dialogue instances, the face to face dialogue text has many functional terms such as pronouns. These are: *re* (1) (we), *ke* (4) (I), *ba* (6) (they), *lo* (7) (you), *lona* (27) (you), *bo* (50) (it/those), *fa* (54) (here/give), *kwano* (58) (here), *nna* (65) (me/sit) and *yo* (70) (this one). Pronouns signal individuals' close interaction with their immediate environment as they point and make reference to where they are.

Other marks of personal interaction are expressed in reactions to what is being said. Such terms include words such as *ee* (2) (yes), *nnyaa* (3) (no), *ehe* (10) (wow/I see), *mm* (16), *mmm* (49) and *ijoo* (24) (interjective of surprise or shock),

Other words that signal interlocutors who are engaging each other are: *ko* (5) (at), *fela* (12) (only), *tlhotse* (17) (spent the day), *hela* (18) (only), *kana* (21) (or), *ga* (22) (of), *jaana* (40) (this way), *eng* (51) (what), *tilwe* (59) (said), *gore* (60) (that), *batla* (64) (want).

As part of the fieldwork, classroom interactions were recorded at junior secondary schools. Since Setswana is used in the teaching of Setswana grammar and literature, only Setswana text from such classes has been recorded, transcribed and added to the corpus. The top 100 keywords from the text are in Table 45.

Table 45: Educational spoken text keywords

N	Keyword	keyness
1	ee	1658.3
2	ke	1384.3
3	letlhalosi	1146.4
4	lebopi	854.85
5	re	823.11
6	sekai	821.37
7	eng	769.09
8	utlwana	726.59
9	felo	679.43
10	kere	590.38
11	ra	496.72
12	dikai	492.56
13	ko	484.48
14	gore	473.02
15	ehe	461.73
16	thito	456.45
17	ga	451.52
18	le	441.43
19	mphang	430.18
20	letlhaodi	427.34
21	raya	422.33
22	hee	402.19
23	popo	376.14
24	ngotlo	333.99
25	kae	323.02
26	mpha	314.13
27	rra	284.99
28	modumo	266.09
29	go	259.6
30	mma	252.9
31	tengwafatso	252.5
32	akere	234.56
33	lerui	233.08
34	letshwaogoka	233.08
35	seyantlo	233.08
36	leemedi	229
37	mphe	223.77
38	o	220.12
39	jaanong	217.85
40	ile	217.2
41	kana	215.64
42	mabopi	213.65
43	masimo	207.92
44	tsamaela	206.38
45	wena	205.7
46	mogatlana	205.44
47	lenyalong	197.21
48	malome	187.78
49	nnyaa	180.37
50	gokelela	176.46
51	fe	175.36
52	ditlhaodi	174.8
53	lesoboki	174.8
54	na	171.63
55	nnya	171.22
56	dirisitse	167.67
57	araba	157.38
58	rinifatso	155.38
59	pirwana	155.38
60	mefuta	149.71
61	mosadi	147.41
62	lone	146.65
63	tlholego	145.3
64	lenyalo	144.02
65	rile	143.03
66	bakang	142.17
67	mothofaditsweng	135.96
68	leamanyi	135.96
69	poufatso	135.96
70	la	132.91
71	leina	131.77
72	bua	129.55
73	itse	123.81
74	sengwe	123.58
75	kwala	122.63
76	utlwe	122.44
77	lengwe	120.26
78	reng	119.77
79	supa	116.
80	lefelo	115.92
81	kedibonye	114.13
82	tsholetsa	113.77
83	tlhaloso	111.59
84	fa	111.32
85	waitse	110.85
86	gago	110.02
87	fela	109.63
88	tilodi	109.33
89	yoo	108.63
90	potso	108.53
91	bona	108.19
92	jang	106.95
93	tsamaetse	105.31
94	gona	104.79
95	lebaka	104.51
96	akanya	103.87
97	dirise	103.82
98	efe	102.48
99	letsogo	101.59

100	tsweleleng	101.57

Since most Setswana classes deal with Setswana linguistics this is reflected in the grammatical labels that are captured in the Table 45 list. These include *letlhalosi* (3) (adverb), *lebopi* (4) (morpheme), *thito* (16) (stem), *letlhaodi* (20) (adjective), *ngotlo* (24) (diminution), *modumo* (28) (sound), *tengwafatso* (31) (palatalisation), *leemedi* (36) (pronoun), *mogatlana* (46) (suffix), *lesoboki* (quantitative), *rinifatso* (lateralisation), *leamanyi* (relative), *mothofatso* (personification), and *poufatso* (labialisation). Setswana classes in general teach Setswana grammar, culture and literature. That is why there are many grammatical terms. Other terms give signal to the giving of instructions found in instruction classes. These include amongst others *sekai/dikai* (6/12) (example/examples), *mphe* (37) (give me), *gokelela* (50) (link/connect), *araba* (57) (answer), *kwala* (75) (write), *supa* (79) (show), *tlhaloso* (83) (explanation), *potso* (90) (question), *akanya* (96) (think), and *efe* (98) (which one?) Setswana cultural terms include *utlwana* (8) (be at peace with), *masimo* (43) (farms), *lenyalong* (47) (concerning a wedding), *malome* (48) (uncle), *pirwana* (59) (black colour of a female sheep), *mosadi* (61) (woman), *bakang* (66) (cause/praise), and *tilodi* (88) (black and white animal colour).

The largest part of the spoken subcorpus is made of Hansard text. The text was scanned from Hansard publications from the Botswana parliament. The most frequent 100 keywords are presented in Table 46.

Table 46: Hansard spoken text keywords

N	Keyword	Keyness
1	gore	20560
2	re	17086
3	ke	10741
4	i	9011.6
5	the	8464.9
6	mr	8217.9
7	of	5400.2
8	to	5328.6
9	leng	4743.7
10	jaanong	4665.1
11	ko	4624.6
12	rraetsho	4421.4
13	hansard	4149.3
14	page	4096.8
15	honourable	3876.3
16	that	3777.9
17	speaker	3648.4
18	is	3512.9
19	and	3342.6
20	bua	3210.8
21	ba	3070.2
22	motsamaisa	2953.9
23	debate	2757.2
24	member	2733.3
25	you	2670.8
26	resumed	2634.3
27	we	2506.8

28	palamente	2392.7
29	motion	2248.4
30	dipuisanyo	2238.5
31	rona	2101.4
32	mme	2091.4
33	bo	2087.6
34	bill	2039.1
35	in	1922.5
36	not	1745.8
37	have	1682.7
38	kana	1673.1
39	are	1672.1
40	this	1670.6
41	it	1664.3
42	ra	1622.6
43	fa	1439.8
44	di	1415
45	march	1411.8
46	teng	1411.1
47	jaana	1410.2
48	tona	1401.4
49	nnyaa	1395.1
50	appropriation	1350.1
51	minister	1337.7
52	yone	1313.4
53	second	1266.1
54	be	1247
55	gone	1209.2
56	reading	1201.5
57	go	1181.5
58	for	1135.4
59	leboga	1109.2
60	motlotlegi	1088.9
61	order	1088
62	se	1085.7
63	on	1084.2
64	draft	1071.7

65	ka	1071
66	point	1054.8
67	gona	1052.3
68	tse	1050.3
69	fela	1029.4
70	ntseng	1018.3
71	development	1015.2
72	gongwe	1013.3
73	buang	955.65
74	privatisation	948.36
75	ga	946.16
76	policy	924.36
77	they	909.84
78	batho	898.65
79	goromente	888.41
80	tsone	874.72
81	rra	861.48
82	ntse	861.1
83	tie	857.75
84	one	850.56
85	head	848.72
86	tuesday	833.77
87	dilo	828.97
88	wednesday	818.69
89	speech	814.07
90	thursday	807.58
91	eleng	798.9
92	clarification	798.86
93	batswana	791.79
94	as	787.79
95	what	781.11
96	there	767.99
97	itse	755.23
98	but	752.49
99	monday	752.22
100	ministry	746.5

The Hansard as an official report of parliamentary speeches has a preponderance of formal parliamentary terminology such as *Mr.* (7), *rraetsho* (13) (Sir), *honourable* (16) or *Motlotlegi* (65), *speaker* (18) or *Motsamaisa* (23) *Dipuisanyo* (31) in Setswana, *member* (25), *palamente* (29) (parliament), *motion* (30), *bill* (35), *Minister* (52), point (71) of (8) order (66), reading (60), draft (69), and *goromente* (86) government. The subcorpus does display common terms in speech such as *gore* (1) (so that), *re* (2) (we), *ke* (2) (I), *the* (5), *of* (7) to (8), *leng* (9) (when), *jaanong* (10) (now), *ko* (12) (at) *that* (16) is (18), it (41) *yone* (52) or *gone* (55) or *gona* (67), in (35), and *not* (36). Other terms like *appropriation* (50), *development* (71),

privatisation (74), and *clarification* (92) indicate the high register which often characterise parliamentary debates. What may be observed is the high occurrence of English terms even in what is Setswana text. There are at least two explanations for this. First, English is an official language in Botswana and educated speakers tend to code switch freely, particularly in official contexts such as parliamentary debates. Second, parliamentary debates make use of specialised terminology which Setswana language has not been developed to handle adequately. Some of the instances of English terminology usages are *bill* (34), *in* (35), *not* (36), *have* (37), *March* (45), *appropriation* (50), *second* (51), *be* (54), *reading* (56), *for* (58), Monday (99), and *ministry* (100).

Spoken subcorpus also comprises a television interview from the Botswana television (Btv) program, *The Eye*. The recorded program was about water conservation in Gaborone in light of the 2004 drought which nearly dried the Gaborone dam which supplies the city of Gaborone with water.

Table 47: Interviews spoken text keywords

N	Keyword	Keyness
1	metsi	2339.9
2	gore	2202.7
3	re	2064.4
4	ke	1664.1
5	ee	1252
6	leng	1249.5
7	eh	1047.5
8	water	955.97
9	ehe	885.45
10	mma	869.78
11	le	677.77
12	ko	386.17
13	utilities	338.61
14	na	302.3
15	eng	292.63
16	bo	272.35
17	rona	271.35
18	dam	261.11
19	letsibogo	258.01
20	mang	223.55
21	be	216.35
22	gone	215.13
23	demand	203.92
24	fem	203.92
25	one	198.55
26	ga	197.75
27	kana	188.73
28	go	186.67
29	so	182.44
30	dintshu	169.93
31	affairs	169.52
32	mme	169.37
33	raya	167.7
34	ra	162.9
35	pipe	161.92
36	fela	161.34
37	nnyaa	156.86
38	dikoko	156.32
39	bua	154.24
40	itse	151.05
41	jaana	150.02
42	lehuma	148.08
43	supply	145.13
44	technology	140.52
45	ka	140.13
46	mo	138.08
47	map	136.66
48	mananeo	133.51
49	fifty	132.19
50	the	130.54
51	jaanong	125.67

52	tweng	123.17
53	matamo	122.19
54	di	122.12
55	waste	119.46
56	litres	118.95
57	letamo	116.44
58	nosa	114.43
59	tlhatswa	112.87
60	pit	106.95
61	nnya	106.69
62	gago	106.55
63	dirisa	106.36
64	femp	101.96
65	menoto	101.96
66	latrine	101.96
67	gaborone	101.49
68	jang	99.687
69	batswana	98.081
70	rre	95.613
71	carrier	94.935
72	nna	94.054
73	dirise	91.167
74	pompa	90.531
75	ceda	89.293
76	ao	89.109

77	tamong	87.031
78	biditswe	86.29
79	tse	86.233
80	lone	85.584
81	maybe	84.964
82	mathata	83.488
83	teng	81.18
84	lona	80.876
85	ba	78.587
86	dilo	76.233
87	tla	76.096
88	tlase	75.679
89	sentle	75.489
90	gape	74.288
91	eight	74.208
92	conservation	74.208
93	tamo	74.208
94	mmitsa	72.787
95	two	72.635
96	kgona	71.154
97	gogwe	71.005
98	metseng	70.544
99	four	70.432
100	corruption	69.994

The subject dealt with in the interviews is clearly revealed by the terms that are key. These include *metsi* (18) (water), (water) *utilities* (13) (the water provider in cities), *dam, letamo, tamo* (18, 57, 93) (dam), *Letsibogo* (19) (the name of a dam), *demand* (23), (water) *affairs* (31) (the water provider in villages), *pipe* (35), *supply* (43), *technology* (44), *waste* (55), *litres* (56), *nosa* (58) (serve water), *tlhatswa* (59) (purify), *mathata* (82) (problems), *conservation* (92).

What stands out as well in this list is a high level of English usage since the subject is technical. These include carrier *water* (8), *utilities* (13), *demand* (23), (71), *maybe* (81), *eight* (91), *conservation* (92), *four* (99), and *corruption* (100).

The Open radio programming subcorpus includes a variety of different radio programs. Amongst these are *Matimela* (a program about lost and found cattle), *Tatediso ya dikgang* (a news program that follows evening news featuring reports from reporters from around the country), *Borukutlhi* (an anti-crime program), and *Molemi-ithute* (an educational program for farmers). The subcorpus is therefore diverse in its coverage since *Tatediso ya dikgang* as a news program covers a variety

of subject matters. Because of the two programs for farmers there are many agricultural terms in Table 48 list.

Table 48: Open radio programming keywords

N	Keyword	Keyness
1	ba	857.19
2	ko	655.83
3	tshipi	440.85
4	tshwailwe	401.26
5	le	318.42
6	boitaolo	258.23
7	re	228.15
8	tlhomagane	206.36
9	di	194.94
10	ya	194.87
11	tlhaka	193.51
12	go	169.04
13	pelesa	168.32
14	serope	162.18
15	mojeng	148.91
16	molemeng	145.57
17	sekolong	145.28
18	baithuti	134.17
19	ke	126.12
20	gore	123.15
21	mo	111.68
22	ka	110.39
23	superintendent	105.07
24	bana	104.62
25	tse	103.41
26	la	98.832
27	kgomo	98.613
28	mathateng	98.369
29	khunwana	88.834
30	mapodisi	85.007
31	sekolo	82.43
32	kgaolong	80.714
33	moroba	77.782
34	batsadi	76.667
35	khamphane	76.652
36	lephaga	76.415
37	maphaga	76.415
38	wa	76.299
39	bao	75.991
40	tlase	73.234
41	nako	72.476
42	godimo	72.133
43	eo	69.95
44	bone	69.666
45	na	68.534
46	batshabi	66.353
47	kwena	65.801
48	bakgweetsi	62.958
49	ngwe	62.165
50	dikgang	62.085
51	Hiv	60.917
52	Te	59.219
53	foods	57.667
54	tsa	57.532
55	selebi	56.715
56	matshwao	56.715
57	aids	52.997
58	bag	52.713
59	isa	52.512
60	tsela	52.492
61	bese	51.04
62	bo	51.04
63	bdp	50.496
64	mapodise	50.307
65	party	47.854
66	kgongwana	47.758
67	tsenya	47.532
68	mme	47.233
69	phati	46.948
70	fitileng	46.59
71	diphologolo	46.304
72	maswabi	46.235
73	khunou	45.386
74	babelaelwa	45.386
75	kopa	44.905
76	kgabaganya	44.854
77	dipalo	44.737
78	matimela	43.918
79	leng	43.098
80	dijo	42.797
81	dira	42.613
82	mokgweetsi	42.456
83	tatediso	42.237
84	rile	41.965
85	neng	41.464
86	lwetse	41.343
87	rre	40.758
88	fa	39.813
89	khampane	39.595

90	theme	39.417
91	mosong	39.327
92	pharakano	38.75
93	tshologa	38.272
94	mokgaoganyi	38.207
95	cosatu	38.207

96	tla	37.933
97	mmuso	37.786
98	dikotsi	37.718
99	jaana	37.281
100	constable	37.272

Because of the subjects handled by the programs, there are terms related to branding of cows such as, *tshipi* (3) (metal used for branding cows), *tshwailwe* (branded), *tlhaka* (11) (branded letter), *serope* (14) (thigh, where cows are branded), *mojeng* (15) and *molemeng* (16) (right and left side; sides on which cows are branded), *lephaga/maphaga* (36/37) (a type of animal ear mark), and *matshwaô* (56) (marks). Other terms refer to the kind and or size of the animal. These include, *pelesa* (13) (heifer), *kgomo* (27) (cow), *moroba* (33) (mid-sized cow), and *kgongwana* (66) (calf). Other terms refer to the colour of the cows. These are *khunwana* (29) (reddish brown on female cows) and *khunou* (73) (reddish brown on male cows).

Other terms point to crime prevention and police work. These include *boitaolo* (6) (rebelliousness), superintendent (23), *mapodisi/mapodise* (30/64) (police officers), *babelaelwa* (74) (suspects), *mokgweetsi/bakgweetsi* (82/48) (driver(s)), *dipalo* (77) (statistics/numbers), *pharakano* (92) (traffic), *dikotsi* (98) (accidents) and constable (100).

Other words are educational. These include *sekolong/sekolo* (17/31) (of school/school), *baithuti* (18) (learners). There are a variety of terms which probably come from different news items. These include *khampane* (35) (company), *nako* (41) (time), *godimo* (42) (above/on top), *batshabi* (46) (refugees), *kwena* (47) (crocodile), Aids (57), bag (58), *tsela* (60) (way/road), *bese* (61) (bus), BDP[26] (63) (the Botswana ruling party), *kgabaganya* (76) (cross), *mosong* (91) (morning), *mmuso* (97) (government).

There is also a considerable use of pronouns such as *ba* (1) (of), *ko* (2) (at), *re* (7) (we), *mo* (21) (in), *tse* (25) (these), *bao* (39) (those), *eo* (43) (that one), *bone* (44) them, *fa* (88) (here). These are common in spoken language.

[26] The Botswana Democratic Party

The religious spoken text is exclusively from the Christian faith. The other's faiths have a very small following nationally (Christian 71.6%, Badimo 6%, other 1.4%, unspecified 0.4%, and none 20.6% (The Republic of Botswana: Central Statistics Office, 2001 census)). The data comprises sermons from churches and funerals and from the radio program, *Sidilega* (be well). The keywords from this data follow in Table 49.

Table 49: Religious spoken text keywords

N	Keyword	Keyness
1	Re	1257.1
2	ke	1125.4
3	modimo	661.24
4	mme	412.19
5	bagaetsho	392.82
6	gago	378.96
7	amen	296.5
8	leboga	293.06
9	le	293.02
10	lefoko	287.22
11	mma	284.13
12	lo	270.29
13	eh	236.18
14	yo	232.77
15	morena	201.33
16	rra	195.2
17	jaanong	185.97
18	gore	181.87
19	tle	177.6
20	tlaa	159.44
21	ga	147.87
22	ka	140.81
23	mo	139.63
24	batsadi	137.04
25	galalelang	136.11
26	bua	130.29
27	bagaetshong	127.95
28	kwano	127.53
29	nne	124.24
30	baruti	123.74
31	rona	114.67
32	be	107.29
33	kagiso	105.6
34	ngwana	105.02
35	fano	104.1
36	bo	103.27
37	jaana	101.1
38	fa	100.6
39	keresete	99.279
40	sefela	98.994
41	tleng	97.905
42	jeso	97.857
43	ntse	93.265
44	robala	93.147
45	mokgatlho	92.883
46	mowa	86.582
47	nna	85.722
48	wena	81.657
49	pholo	80.714
50	teng	80.031
51	itse	78.64
52	raya	77.168
53	senatla	76.036
54	rraetsho	75.592
55	ralekgotla	72.16
56	burial	71.876
57	seabi	71.876
58	kwaletswe	67.288
59	moruti	66.894
60	christ	65.859
61	tsena	64.864
62	phuthego	64.433
63	malebogo	63.094
64	tlhodilwe	62.365
65	ulululuuuu	61.607
66	hodisa	61.607
67	papa	60.058
68	sebui	59.819
69	baebele	59.764
70	nnyaa	59.736
71	wa	57.191
72	ko	56.166
73	kalo	55.268

74	ene	55.095
75	nae	54.952
76	tsaa	54.844
77	ha	53.71
78	tsile	53.587
79	buang	53.477
80	dumedisa	51.689
81	fela	51.471
82	rapeleng	51.339
83	tshegofatso	51.243
84	gonne	50.687
85	ra	50.603
86	matlhogonolo	50.052
87	three	50.052
88	go	49.754
89	dithaba	48.959
90	kgotleng	48.245
91	ira	48.194
92	mokgatlhong	47.949
93	lapeng	47.295
94	rata	47.028
95	tlaabo	46.927
96	khwaere	46.438
97	tshikinyega	45.944
98	na	45.591
99	masego	45.377
100	mmoloki	44.531

Most of the keywords are from the Christian religion. This is evident in the list with words such as: *Modimo* (3) (God), Amen (7), *morena* (15) (lord), *Keresete* (39) (Christ), *sefela* (40) (hymn), *Jeso* (42) (Jesus), *Christ* (60) *phuthego* (62) (congregation/a gathering), *Baebele* (69) (Bible), *rapeleng* (82) (pray), *tshegofatso* (83), *matlhogonolo* (86) and *masego* (99) (blessing(s)), and *mmoloki* (100) (saviour). From the funeral service words which stand out are mokgatlho (45) ((burial) society), burial (56), *sebui* (68) (a speaker), *kgotleng* (90) (the traditional meeting place where a meeting is held before burial), *ralekgotla* (55) (headman) *mokgatlhong* (92) (society) and *khwaere* (96) (choir). The funeral service as a formal gathering is characterised by words of respect such as *bagaetsho* (5) or *bagaetshong* (27) (fellow citizens), *Mme* (4) (Mrs), *Mma* (11) (mother of or Mrs.), *rra* (16) (Sir) and *rraetsho* (54) (Sir). The *Sidilega* program wishes the sick persons good health through the use of biblical scriptures. It therefore speaks of *pholo* (49) (healing) and *hodisa* (heal) (66).

The spoken subcorpus also comprises football commentary and other radio programs covering other types of sports. The keywords of this text are given in Table 50.

Table 50: Sport spoken text keywords

N	Keyword	Keyness
1	kgwele	1162.3
2	Ko	964.36
3	Ke	942.04
4	motshameko	562.01
5	team	559.96
6	setlhopha	461.49
7	mme	457.78
8	small	440.84
9	lebelela	436.12
10	mokatisi	421.29
11	coach	378.14

12	tshamekela	371.11
13	tsaya	353.65
14	e	344.49
15	ditaola	328.2
16	gore	325.92
17	le	308.5
18	league	305.12
19	tsewa	297.16
20	boy	295.42
21	viola	287.69
22	eo	286.04
23	tshameka	254.16
24	fale	251.46
25	tla	242.95
26	softball	239.85
27	ya	235.35
28	kadisa	226.86
29	pitcher	226.86
30	ka	224.71
31	lebelele	222.37
32	metshameko	220.67
33	pikati	202.82
34	wa	191.53
35	re	186.58
36	matius	181.02
37	ketshabile	181.02
38	lefela	179.72
39	gabaitsane	179.59
40	bakale	179.59
41	tatlhelo	179.59
42	mosimane	179.07
43	jwaneng	176.1
44	bona	175.03
45	fitileng	174.3
46	nako	174.23
47	tshutshu	170.14
48	na	168.23
49	ee	164.55
50	leng	162.25
51	jono	162.12
52	spears	160.69
53	setlhopheng	151.92
54	police	151.53
55	kgantele	148.74
56	kamoso	148.03

57	ntse	138.29
58	yone	137.87
59	tlosa	136.74
60	motshekgwa	134.32
61	wells	132.33
62	ecco	125.42
63	yole	125.03
64	masie	125
65	angels	125
66	blue	124.93
67	bokhutlo	122.16
68	tswela	117.96
69	netball	116.77
70	gunners	116.67
71	bokhutlong	114.06
72	tle	113.54
73	tournament	113.42
74	morapedi	113.21
75	satmos	112.52
76	go	111.02
77	tsatsing	108.1
78	gone	107.29
79	beke	106.52
80	chiko	106.39
81	player	106.39
82	kenny	106.39
83	fifa	105.4
84	okay	103.97
85	tshobega	103.97
86	motshamekong	101.73
87	lebe	101.18
88	duncan	97.821
89	modirelabangwe	97.105
90	themba	96.628
91	mabogo	96.217
92	bone	95.895
93	pitch	94.519
94	stopo	94.519
95	catcher	94.519
96	molokwane	94.519
97	shephi	94.519
98	teng	94.378
99	bokatisi	91.562
100	friday	89.968

The results of Table 50 are characterised by sport terms that refer to a variety of **games** such as *kgwele* (1) (ball or football), *softball* (26), and *netball* (69). Sport spoken text comprises football commentaries and interviews of sport personalities. Some of the terms that come up amongst the top 100 include amongst others names of **footballers**: *Ditaola* (15), *Viola* (21),

Kadisa (28), *Pikati* (33), *Matius* (36), *Ketshabile* (37), *Bakale* (40), *Motshekgwa* (60) and many others. There are also **names of teams**: *Spears* (52), *Police* (54), *Wells* (61), *Ecco* (62), *Blue Angels* (66/65), *Gunners* (70), and *Satmos* (75). **Nouns and verbs common in sport** also rank high, amongst these being *kgwele* (1) (ball), *motshameko/metshameko* (4/32) (sport/game(s)), *setlhopha* (6) (team), *lebelela* (9) (watch), *mokatisi* (10) (coach), *coach* (11), *tshamekela* (12) (play for), *tsaya* (13) (take), *league* (18), *tshameka* (23), *pitcher* (29), *lebelela* (31) (watch), *lefela* (38) (zero), *tatlhelo* (41) (throw in), *nako* (46) (time), *tournament* (73), *player* (81). A variety of **sports** are represented as well amongst these being *softball* (26) and *netball* (69).

From the list above, there is an interesting use of certain words whose use appears to be unique to sport. Of note is the word *lebelela* which means 'to watch or observe'. One would expect the use of *lebelela* in the area of sport to have spectators as the subject and the game as the object of the verb. This however is not the case. The structure that we get in the concordance lines is that of *O a lebelela* (He is watching) followed either by *ka kgwele* (with the ball) or *ke Kenny Ramco (the name of a player).*

	Concordance
30	o e tshamekela ko go Pikati. Pikati o a lebelela ka kgwele eo, o e tshamekela ko go
31	kgwele eo a kgorelediwa ke Thobega. O a lebelela o sireleditse bontle fale Thobega
32	e ya go tsewa fale ke Duncan. Duncan o a lebelela ka kgwele eo a kgorelediwa ke
33	Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny, Kenny o bona fale
34	o a lebelela a kgwele eo ya gagwe o a lebelela ke Ramco. A re o feta ka Bakale e
35	ko mosimane yo Molokwane, Ramco o a lebelela a kgwele eo ya gagwe o a lebelela ke
36	fitlha fa. E ya go tsewa ke Pikati, Pikati o a lebelela ka kgwele eo tshamekela ko go
37	yo Kenny Ramodisa, Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny,
38	kgwele ya tsewa ke Betsho, Betsho o a lebelela ka kgwele eo o bona Ditaola, Ditaola
39	ba ya go e tsaa ba e tlosa kgwele eo. O a lebelela ke Themba Ketshabile ka kgwele eo.
40	o e tshamekela ko go Pikati. Pikati o a lebelela ka kgwele eo, o e tshamekela ko go
41	kgwele eo a kgorelediwa ke Thobega. O a lebelela o sireleditse bontle fale Thobega
42	e ya go tsewa fale ke Duncan. Duncan o a lebelela ka kgwele eo a kgorelediwa ke
43	Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny, Kenny o bona fale
44	o a lebelela a kgwele eo ya gagwe o a lebelela ke Ramco. A re o feta ka Bakale e
45	ko mosimane yo Molokwane, Ramco o a lebelela a kgwele eo ya gagwe o a lebelela ke
	fitlha fa. E ya go tsewa ke Pikati, Pikati o a lebelela ka kgwele eo tshamekela ko go
	yo Kenny Ramodisa, Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny,

In these instances when a player takes a football and looks to where he could pass it, the very act of searching for an unmarked player is expressed by the verb *lebelela*. This use of *lebelela*

is unique to sport since the common use of *lebelela* is watch or watch over something.

Another word similar to *lebelela* which is unique to sport is the word *tsaya* (take). We first present its concordance lines

110	lhopha sele sa Police11. ba ya go e tsaya ba e tlosa kgwele eo. Ba ya
111	a re lebelele kgwele eo, ba ya go e tsaya ba e tlhoma. motsotso wa bom
112	wela kwa ntle kgwele eo. Ba ya go e tsaya ba e tlhoma ba etlosa go tlo
113	ousand and four . Kgwele ba ya go e tsaya ba e tlosa, motshameko e le
114	etsa go tswa kwa morago. Ba ya go e tsaya ba e emeletsa go tswa ka kwa
115	se kgobalo epe e masisi. Ba ya o e tsaya ba e tlosa setlhopa sa BDF11
116	osa e le tatlhelo BDF11. Ba ya go e tsaya ba e tlosa e le tatlhelo. Mo
117	ogo ya ga Oliver Pikati. Ba ya go e tsaya ba e tlhoma e le goal kick k
118	le kgwele e ntle. Kgwele ba ya go e tsaya ba e kolopa ba e emeletsa go
119	ousand and four . Kgwele ba ya go e tsaya ba e tlosa, motshameko e le
120	lopa Police11 kgwele eo. Ba ya go e tsaya ba e kolopa e kolopiwa fa le
121	a fela kontle kgwele eo. Ba ya go e tsaya ba e kolopa e le tatlhelo. B

To take something implies the use of hands. However in this instance this is not the case. It simply means being in possession of a ball.

Some of the words that collocate with *tsaya* are also unique to the genre of sports especially when referring to football. The words are *tlosa* (remove), *kolopa* (throw at), *tlhoma* (fix on the ground), and *emeletsa* (raise up/lift up). *Kolopa* implies the use of hands to throw something at someone or something, however in this context it refers to kicking a ball into the air. *Emeletsa* is to raise something upright or on its feet. In the sports genre however it refers to setting a ball into flight. The use of these words indicates that words function differently in different contexts and their treatment in dictionaries need to reflect the different contexts in which they are used. For instance the treatment of the *tsaya* in Matumo (1993) may be improved by extracting concordance lines from a corpus and identifying collocates. *Tsaya* is entered in Matumo (1993: 426) thus:

> **tsaya** V.S. SIMP., take; take a wife; marry.

This entry can be improved this way:

> **tsaya** *v.* **1.** take with hands **2.** follow a path, choose a direction **3.** take a wife; marry **4.** be in possession of ◘ **tsaya botshelo**: take a life ◘ **tsaya dinopolo**: spy on someone. ◘ **tsaya ditaelo**: take orders. ◘ **tsaya ka motlhala**: follow. ◘ **tsaya dipilisi**: swallow pills. ◘ **tsaya dinopolo**: collects secrets ◘ **tsaya karolo:** take part. ◘ **tsaya ka letsogo**

la molema: illtreat; discriminate against. ◘ **tsaya kgakololo**: take advice. **tsaya kgato**: take a step. ◘ **tsaya lobaka**: take a long time. ◘ **tsaya mongwe/sengwe motlhofo**: undermine someone or something. ◘ **tsaya puso**: take over government. ◘ **tsaya phekelo e sele**: take a turn for the worst. ◘ **tsaya tshwetso**: take a decision. ◘ **tsaya nako**: take time. ◘ **tsaya motlhala**: copy an example from someone. ◘ **tsaya mosadi**: take a wife; marry. ◘ **tsaya mogote**: measure temperature. ◘ **tsaya matsapa**: put an effort. ◘ **tsaya tsia**: take someone or something seriously. ◘ **tsaya malatsi**: go on leave. ◘ **tsaya malebela**: copy something good. ◘ **tsaya maikarabelo**: take responsibility. ◘ **tsaya loeto**: take a trip. ◘ **tsaya maemo**: occupy a position. ◘ **tsaya setshwantsho**: take a picture. ◘ **tsaya sekgele**: win an award. ◘ **tsaya sebaka**: take time.

6.3 Conclusion to keyword analysis

In the preceding pages we have calculated keywords for Science and Technology, Politics, Poetry, Plays, Grammar, Arts and Culture, Religious and Spoken texts. We have presented the top 100 keywords from each genre or text type and shown that every text type contributes unique words. The findings support the position that for a corpus to represent the general language, it must be designed in such a way as to include a variety of text types from the language. The finding has been supported by keyword analysis which has revealed that the different text types generate different keywords that are particular to them.

The recognition that different text types contribute different words, should influence lexicographers compiling dictionaries on the basis of corpus evidence to pay particular attention to corpus design to ensure the broadest coverage possible of text types in a corpus. This is since the quality of retrieved information for lexicographic purposes depends on the information input at the stage of corpus construction.

Additionally, lexicographers could harness the power of keyword analysis and mark dictionary entries and senses on the basis of word variability. Many English dictionaries consistently mark frequencies (Kilgarriff, 1997; Summers, 1995); however much can be achieved by marking words or senses which rank high in a particular genre or text type. The challenge with raw frequency lists generated from a whole corpus is that they can push words which are high on the frequency analysis of a specific genre down on the frequency list of the whole corpus. The solution lies in an analysis similar to the ones conducted in this chapter which are genre based. As a result of keyword analysis in this chapter for instance, the words in Table 51 could

therefore be entered in a dictionary and marked SPORT to indicate that they rank high in the keyword frequency analysis of sports terms.

Table 51: Possible SPORT candidates

English	English
kgwele	ball
motshameko	game
setlhopha	team
lebelela	watch/look
mokatisi	coach
tshamekela	play at
liki	league
tshameka	play
softball	softball
pitcher	pitcher
metshameko	games
tatlhelo	throw in

We illustrate the labelling with two dictionary entries *kgwele* and *setlhopha* from Matumo (1993).

> **kgwele** N. CL. 9N-, SING., any round object; commonly used to refer to a football.
>
> **setlhôpha** N. CL. 7 *se-,* SING OF *ditlhôpha*, a group; a company of people; a drove of animals.

The two entries could be improved with the SPORT label this way.

> **kgwele** *n*. [SPORT] **1.** a football. **2.** a ball.
>
> **setlhôpha** *n*. **1.** [SPORT] a sports team **2.** a group.

Marking entries as suggested will aid users and language learners in identifying the genre in which the word functions even before reading an illustrative sentence in an entry. Lexicographers could therefore device labels such as RELIGION, MUSIC, GRAMMAR, ARTS, NEWS, POLITICS, SCIENCE or LAW to make the dictionary more informative and user friendly.

In Chapter 7 we measure lexical density across text types at comparable token points. It will

be established whether at comparable token points text types vary in lexical density and contribute different words. The diversity of lexical richness found in genres and domains is relevant for dictionary compilation since as argued before dictionaries should aim to be broad in their coverage of a language's lexicon.

Chapter 7

Type/token measures of corpus chunks

7.1 Type/token measures

This chapter measures the degree to which with every additional 10,000 tokens the number of word types grows. Type/token ratio measures lexical richness and determines lexical closure in a text or corpus. If the number of types grows with the addition of every 10,000 tokens it will show that a text has not reached lexical closure. If on the other hand the types do not grow, it will signal lexical closure. What is investigated is the degree to which types grow at comparable points since we seek to determine the lexical richness of different text types at comparable points. The question has been suggested differently by Kjellmer:

> Another method of measuring the density of a text type could be to try and answer the question: How many words (types) has the writer introduced into his text after 100 running words (tokens), how many after 200, etc? The more types he has introduced, the more varied his style is likely to be (Kjellmer, 1994: 117).

The aim is therefore to investigate how types grow at comparable token points in different text types. The purpose of the experiments is to establish whether text types vary in lexical density. The diversity of lexical richness found in genres and domains is crucial for the application of lexicography since a dictionary that aims to capture the language variability will be enriched by a corpus comprising texts from diverse sources.

The results of type measures experiments at comparable points are then plotted in a graph to graphically reveal the text types with both high and low text type growth. The experiments are significant in that they measure word types in different text types at similar numerical intervals making it possible to make useful comparisons between text types.

Statistical studies of vocabulary usually report the ratio between types and tokens for a given sample of text (Baayen, 2001). However such statistics are rarely informative since as more word repetitions occur, the type-token ratio falls regardless of the text studied. The TTR is bound to decline towards zero as tokens increase.

Because of such a phenomenon Youmans argues that:

> …this ratio cannot distinguish any text (or any author) from any other. It is not type-token ratios that are significant, *but only the rate at which they decline.* …Type-token ratios are meaningless, then, unless we also specify the number of tokens used in computing them… But this makes it pointless to compute a ratio at all, since this ratio provides no more information than the raw data do… That is, we can compare the number of types directly rather than the type-token ratios and the ratio between these two pairs of statistics is necessarily the same … it is preferable to plot the number of types in a passage directly against the number of tokens, rather than type-token ratios (Youmans, 1990: 588, italics mine).

Following Youmans, in this study the number of types is plotted directly against the number of tokens of various text types at comparable points. The corpus text types are divided into fifty 10,000 tokens chunks. That is, although other text types have many tokens that could exceed fifty 10,000 word chunks, for these experiments we use only 50 chunks (500,000 tokens). Some text types such as Science and Business have fewer tokens comparatively. As the smallest text types they each has 140,000 tokens and 100,000 tokens respectively, therefore their tokens fail to reach the 500,000 tokens measurement. Although they are smaller comparatively, they are still large enough to be used for useful comparisons.

Measuring series of text chunks at comparable tokens for word types is however sensitive to the order in which the texts (i.e. 10,000 word corpus chunks) are ordered. This is problematic since every experiment repetition is likely to give different results depending on which one of the 10,000 token-chunks was analysed first.

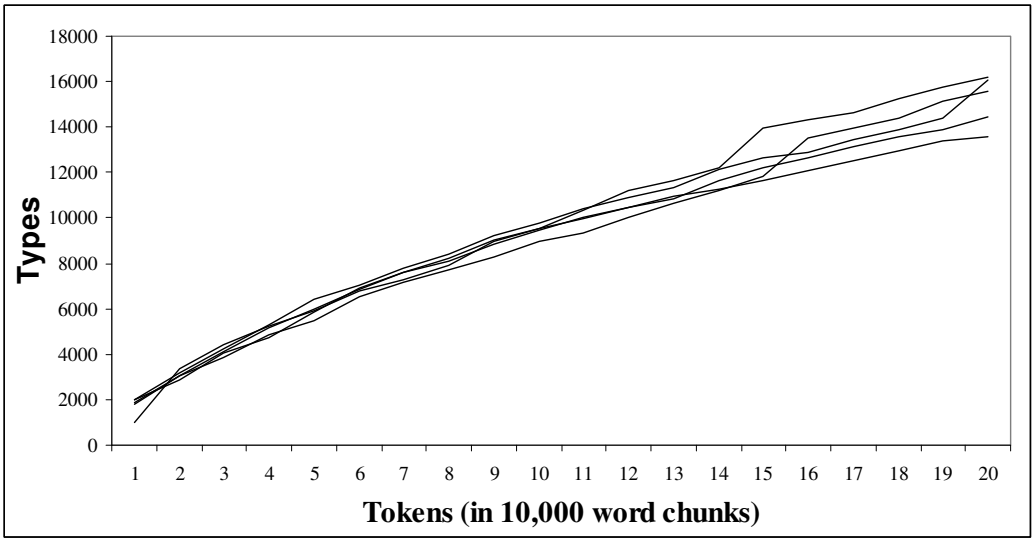
We illustrate this matter below with the five experiment measurements of types from newspaper at 10,000 token intervals up to 200,000 tokens.

Table 52: Newspaper types at 10,000 word tokens intervals

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5
10000	1986	1889	1835	1986	1021
20000	3165	3025	3063	2893	3348
30000	4232	3873	4119	4069	4394
40000	5311	4865	5151	4728	5249
50000	6398	5454	5953	5872	5913
60000	7049	6554	6833	6927	6759
70000	7759	7161	7585	7616	7301
80000	8382	7695	8110	8244	7932
90000	9194	8257	8840	9050	8955
100000	9790	8954	9476	9541	9548
110000	10405	9360	10011	9940	10354
120000	10909	10024	10443	10437	11219
130000	11366	10637	10948	10860	11625
140000	12166	11193	11253	11647	12227
150000	12629	11857	11651	12236	13934
160000	12863	13497	12075	12632	14319
170000	13443	13923	12541	13124	14647
180000	13878	14411	12943	13558	15288
190000	14401	15119	13364	13883	15786
200000	16071	15585	13572	14480	16206

Although Table 52 shows measurements of types from Newspaper text type, whenever an experiment is repeated with a different 10,000 token chunk, this results with a different word type counts. There is a variability of types of the same size from the same text type at comparable token points. For instance, at 200,000 token points there is the following variability of types: 16071, 15585, 13572, 14480, and 16206. This is apparent particularly in Figure 10.

Figure 10: Newspaper types at 10,000 word tokens intervals



The variability of types results from the fact that different 10,000 token chunks are measured at the different token points.

7.1.1 The Mean calculation

To resolve the bias of sequence, the 10,000 token-chunks are randomised for every measurement taken and the experiment iterated five times. The type measurements are taken at every 10,000 token intervals up to 500,000 tokens, repeated five times and an average computed. This is so that we could make comparisons between text types using a single mark or an average that summarises the results i.e. gives an average of types at every 10,000 tokens interval. We therefore compute the measure of central tendency, for which we have chosen the mean. We calculate the mean of the scores using the following formula:

$$\bar{x} = \frac{\sum x}{n}$$

$\bar{x}$ is used for the sample mean; $\sum$ means "the sum of"; x indicates a score and n is used for the number of sample scores. The symbols $\sum x$ means 'add up all the scores'.
The mean is therefore calculated by adding all the scores and dividing their total by their sample size.

The Table 52 scores can therefore be rendered as a table of means that summarises the scores as follows in Table 53:

Table 53: A table of means for Newspaper types

Tokens	Mean
10000	1743.4
20000	3098.8
30000	4137.4
40000	5060.8
50000	5918
60000	6824.4
70000	7484.4

Tokens	Mean
80000	8072.6
90000	8859.2
100000	9461.8
110000	10014
120000	10606.4
130000	11087.2
140000	11697.2

Tokens	Mean
150000	12461.4
160000	13077.2
170000	13535.6
180000	14015.6
190000	14510.6
200000	15182.8

7.1.2 Confidence Interval (CI) calculation

Rather than choosing a single value for the population mean, we can specify a range of values within which we are confident that the value lies (Hinton, 2004: 69). We choose a level of confidence, usually 95% or 99% level of confidence, and then work out the range of values. A level of confidence is the probability that the interval estimate contains the population parameter (Larson and Farber, 2006: 281). If we choose the 95% confidence interval, we are saying that if we worked out the confidence interval for 100 different samples from a population the 95% of those confidence intervals would contain the population mean.

To calculate the confidence interval within which a sample mean lies, we need to know the critical value, standard deviation and the sample size. For our experiments we use 95% confidence interval level as our critical value.

7.1.3 Standard deviation

The standard deviation is a measure of how widely values (raw scores) are dispersed from their mean. We calculate the sample standard deviation using the following formula:

$$s = \sqrt{\frac{\sum (x - \bar{x})2}{n-1}}$$

The lower case *s* represents standard deviation. $\sum(x - \bar{x})^2$ means 'subtract the mean from each raw score to find the deviation score, then square each deviation score and add them all up'. $n - 1$ is what is known as the nonbiased method based on degrees of freedoms (*df*) – the total number of samples minus one. Degrees of freedom concern the scores that contain new information. Pagano (2001: 292) defines degrees of freedom thus: "The degrees of freedom (df) for any statistic is the number of scores that are free to vary in calculating that statistic." There are *N* degrees of freedom associated with the mean since for any set of scores *N* is given. As we have calculated the sample mean from the sample scores we have used up some of the information in

the scores. The number of scores with new information, the degrees of information, is $n - 1$ (Hinton, 2004: 52). We give the example below of a set of scores, their means and the calculation of their standard deviations.

Table 54: Newspaper type scores with mean and standard deviation scores

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5	Mean	SD
10000	1986	1889	1835	1986	1021	1743.4	409.0114
20000	3165	3025	3063	2893	3348	3098.8	169.9741
30000	4232	3873	4119	4069	4394	4137.4	193.4665
40000	5311	4865	5151	4728	5249	5060.8	252.6108
50000	6398	5454	5953	5872	5913	5918	335.0604
60000	7049	6554	6833	6927	6759	6824.4	186.0371
70000	7759	7161	7585	7616	7301	7484.4	245.493
80000	8382	7695	8110	8244	7932	8072.6	268.7263
90000	9194	8257	8840	9050	8955	8859.2	360.7932
100000	9790	8954	9476	9541	9548	9461.8	308.0101
110000	10405	9360	10011	9940	10354	10014	418.8323
120000	10909	10024	10443	10437	11219	10606.4	464.0666
130000	11366	10637	10948	10860	11625	11087.2	400.1983
140000	12166	11193	11253	11647	12227	11697.2	488.4959
150000	12629	11857	11651	12236	13934	12461.4	904.087
160000	12863	13497	12075	12632	14319	13077.2	861.2184
170000	13443	13923	12541	13124	14647	13535.6	798.8284
180000	13878	14411	12943	13558	15288	14015.6	887.9957
190000	14401	15119	13364	13883	15786	14510.6	964.0256
200000	16071	15585	13572	14480	16206	15182.8	1127.631

Having made calculations of standard deviation and determined to use 95% confidence interval, our sample size is at each 10,000 word-token interval up to 500,000 and we can calculate the confidence interval (CI) for the mean.

What we calculate are the upper and lower limits for the 95% confidence interval. We achieve these by calculating the area under the standard normal curve that equals 95%. The value for this area is ± 1.96. This implies that 95% of the area under the standard normal curve falls within 1.96 standard deviations of the mean. The confidence interval is therefore:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

We then calculate the left and right endpoints (or the upper and lower limits for the

confidence interval) and form the confidence interval this way (Larson and Farber, 2006: 297):

Left endpoint: $\bar{x}$ - E

Right endpoint: $\bar{x} + E$

Interval: $\bar{x} - E < \mu < \bar{x} + E$

$\bar{x}$ is the sample mean, E is the margin of error and μ is population mean. Below are the results of the calculation of the confidence interval.

Table 55: Newspaper type scores with mean, critical value, standard deviation and confidence interval scores

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5	Mean	CV	SD	CI	LOW	UPPER
10000	1986	1889	1835	1986	1021	1743.4	0.05	409.0	8.0	1735.4	1751.4
20000	3165	3025	3063	2893	3348	3098.8	0.05	170.0	2.4	3096.4	3101.2
30000	4232	3873	4119	4069	4394	4137.4	0.05	193.5	2.2	4135.2	4139.6
40000	5311	4865	5151	4728	5249	5060.8	0.05	252.6	2.5	5058.3	5063.3
50000	6398	5454	5953	5872	5913	5918	0.05	335.1	2.9	5915.1	5920.9
60000	7049	6554	6833	6927	6759	6824.4	0.05	186.0	1.5	6822.9	6825.9
70000	7759	7161	7585	7616	7301	7484.4	0.05	245.5	1.8	7482.6	7486.2
80000	8382	7695	8110	8244	7932	8072.6	0.05	268.7	1.9	8070.7	8074.5
90000	9194	8257	8840	9050	8955	8859.2	0.05	360.8	2.4	8856.8	8861.6
100000	9790	8954	9476	9541	9548	9461.8	0.05	308.0	1.9	9459.9	9463.7
110000	10405	9360	10011	9940	10354	10014	0.05	418.8	2.5	10011.5	10016.5
120000	10909	10024	10443	10437	11219	10606.4	0.05	464.1	2.6	10603.8	10609.0
130000	11366	10637	10948	10860	11625	11087.2	0.05	400.2	2.2	11085.0	11089.4
140000	12166	11193	11253	11647	12227	11697.2	0.05	488.5	2.6	11694.6	11699.8
150000	12629	11857	11651	12236	13934	12461.4	0.05	904.1	4.6	12456.8	12466.0
160000	12863	13497	12075	12632	14319	13077.2	0.05	861.2	4.2	13073.0	13081.4
170000	13443	13923	12541	13124	14647	13535.6	0.05	798.8	3.8	13531.8	13539.4
180000	13878	14411	12943	13558	15288	14015.6	0.05	888.0	4.1	14011.5	14019.7
190000	14401	15119	13364	13883	15786	14510.6	0.05	964.0	4.3	14506.3	14514.9
200000	16071	15585	13572	14480	16206	15182.8	0.05	1127.6	4.9	15177.9	15187.7

Table 55 shows the 10,000 token interval iterations of types-counts. They are followed by the mean calculations of the five iterations. CV stands for the critical value which is at 5% or 0.05. SD stands for the standard deviation which is followed by the confidence interval (CI) calculation results and the upper and lower confidence interval limits.

With a 95% confidence interval we are saying that if we worked out the confidence interval for 100 different samples from the newspaper section of the Setswana corpus, the 95% of those confidence intervals would contain the population mean. For instance at 200,000 token-population that interval is between 15177.9 and 15187.7. The confidence interval calculations are preferable since they show the confidence intervals which contain the population mean.

The rest of the experiments in tables, henceforth give scores as means calculated from five randomised iterations.

7.2 Text divisions for experiments

For our experiments we have divided the Setswana corpus into the following major text types from the written section of the corpus and that of the spoken subcorpus.

First we discuss the written part of the corpus:

Table 56: Written subcorpus text types

1. Poetry
2. Grammar
3. Chat-site
4. Plays
5. Prose
6. Science
7. Politics
8. Business
9. Religious
10. Newspaper

Miscellaneous text has been left out from experiments since it comprises text from different sources and it is not expected to offer useful information for text type comparison.

The spoken subcorpus has been divided into two major parts:

1. Hansard

2. Call-in, interview and open-radio programming treated as a single unit.

Religious and sport text have been left out as too small for meaningful comparisons.

Further, 50 samples of 10,000 token-chunks were sampled from different text types and combined into what could be termed a single created "text type". We achieved this by randomly dividing the 12 text types into three groups with each having four different text types and sampled text from each text type randomly. We labelled these groups using the initial three letters of each text type in the group, thus: POEGRACHAPLA (Poetry, Grammar, Chat-site and Plays), PRONEWHANCAL (Prose, Newspaper Hansard & Call-in) and SCIPOLBUSREL (Science, Politics, Business and Religious). These divisions are given in Table 57.

The aim of the experiment is to determine the results of comparing subcorpora containing unrelated text with equal-sized subcorpora containing text from a single genre. We measure TTR at comparable points for both texts. The claim is not that {Science, politics, business & religious}, {prose, newspaper, Hansard & call-in} and {poetry, grammar, chat-site & plays} groupings are related in any linguistic way – rather the claim is to the contrary – that they are unrelated and each contribute some distinct types. Combining text from a variety of sources therefore (as one might indeed do in corpus compilation) we hope should give a higher TTR at comparable points compared to that of distinct subcorpus measures.

Table 57: Three divisions of text types

A (POEGRACHAPLA)	
Poetry	13
Grammar	13
Chat-site	12
Plays	12

B (PRONEWHANCAL)	
Prose	13
Newspaper	13
Hansard	12
Call-in	12

C (SCIPOLBUSREL)	
Science	13
Politics	13
Business	12
Religious	12

POEGRACHAPLA, PRONEWHANCAL and SCIPOLBUSREL have 500,000 tokens each. The 500,000 tokens for each newly grouped "text type" was achieved by sampling 13 x 10,000 from two text types and 12 x 10,000 from the remaining two text types to get a total of 500,000 tokens.

This brings to 15 the total number of texts measured and compared.

Table 58: Fifteen major corpus text types

1. Poetry
2. Grammar
3. Chat-site
4. Plays
5. Prose
6. Newspaper
7. Hansard
8. Call-in
9. Science
10. Politics
11. Business
12. Religious
13. POEGRACHAPLA
14. PRONEWHANCAL
15. SCIPOLBUSREL

These three (POEGRACHAPLA, PRONEWHANCAL and SCIPOLBUSREL) have been compiled to test two things: whether the combination of chunks from a variety of text types results in a higher types count at each 10,000 tokens interval compared to a count from a single text type. Second, using the whole Setswana corpus' most frequent 100 words as a standard against which to compare, we generate a frequency list for each of the 15 text types listed above and compare each of their 100 most frequent words against the most frequent 100 words of the whole corpus. Frequency lists present an attractive way of looking at text for statistical analysis. Kilgarriff (1997a: 233) offers at least three advantages to using frequency lists:

i. When a text or corpus is represented as a frequency list, much information is lost, but the tradeoff is an object that is susceptible to statistical processing.
ii. An advantage of using frequency lists is that there is so much data: two corpora can be compared in respect of thousands of data points (e.g., words).
iii. Word frequency lists are cheap and easy to generate.

The frequency lists will therefore be used to compare text types. The assumption is that lists drawn from texts compiled from a variety of text types will be similar to the

one drawn from the entire Setswana corpus, while the list drawn from a single text type is expected to be less similar.

First, we give the results of types at 10,000 token intervals starting with POEGRACHAPLA, Poetry, Grammar, Chat-site and Plays text types.

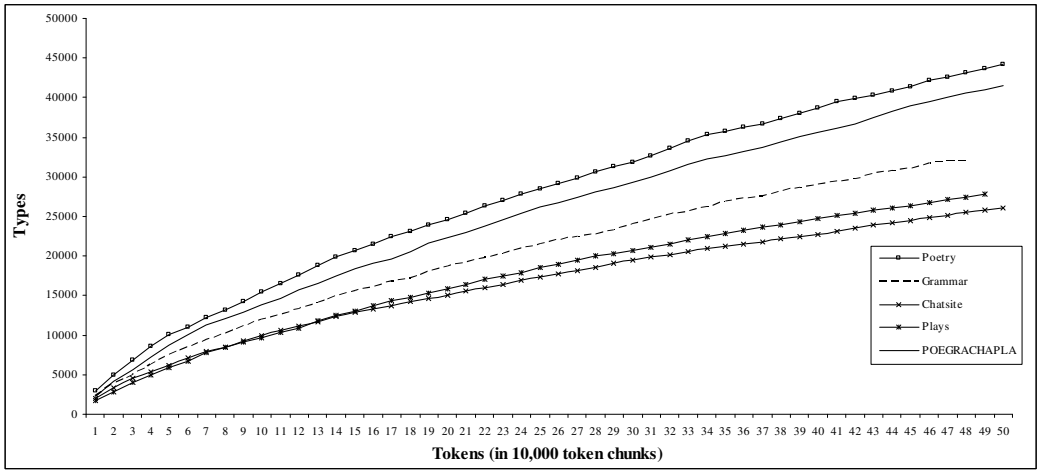
Table 59: Poetry, Grammar, Chat-site, Plays, POEGRACHAPLA text types

Tokens	Poetry	Grammar	Chat-site	Plays	POEGRACHAPLA
10000	2896.6	2274.6	1975.2	1680.2	2223.6
20000	5018	3873.4	3409	2831	4119
30000	6877.4	5022	4528.6	3995	5658
40000	8585.4	6293.8	5323	5015	7210.2
50000	10049.6	7568.2	6121.4	5970.6	8775.6
60000	11057.2	8465	7095.4	6772.2	10127.4
70000	12200.2	9383.6	7882.6	7812.6	11224.6
80000	13222.8	10254.6	8530	8404.2	12120
90000	14241	11157.8	9258.4	9082.6	12884
100000	15494.8	11944.6	9970.6	9683.8	13845.6
110000	16566	12634.2	10587	10321.8	14659.2
120000	17591.4	13292.4	11171	10898.6	15769
130000	18828.4	14085.2	11759.8	11808.8	16595.6
140000	19862	14977.8	12343.8	12475	17528
150000	20677.6	15603.4	12881	13047	18435.8
160000	21509.8	16153.4	13276.4	13762	19109
170000	22399.6	16744	13765	14322.2	19671.6
180000	23134.4	17228	14284.2	14848	20560
190000	23862.8	18022.6	14693.6	15381.8	21645.4
200000	24529.6	18680	15120	15881.2	22354
210000	25341	19275	15634.8	16351.2	22998.8
220000	26308.4	19691.6	16013.2	17004	23764.4
230000	27073.6	20358	16411.2	17502	24568.6
240000	27783.4	21004.4	16959.8	17941.2	25446.8
250000	28517.4	21570.8	17372.2	18487	26267.2
260000	29222.8	22085.2	17808.2	19008	26791.8
270000	29877.6	22474.4	18165.2	19455.6	27434.2
280000	30659	22735.4	18601.8	19974.6	28088.6
290000	31335.6	23263.2	19093.6	20360.4	28621.2
300000	31914.8	24012.4	19426.2	20738.4	29314
310000	32637.2	24644.8	19852.6	21160.4	29939.2
320000	33578.8	25229.8	20224.4	21552.6	30719
330000	34594.2	25724.2	20562.8	22084	31596.2
340000	35307.6	26272.8	20916	22451.4	32216.8
350000	35800.2	26864.4	21214.6	22828.2	32710.2
360000	36300.2	27337	21474.8	23246.6	33254.4
370000	36699.8	27609	21791.6	23605	33800.4
380000	37341.2	28191.8	22111.6	23966.8	34428.8
390000	37977.2	28677	22482.2	24347.8	35037.6
400000	38758	29090.2	22781	24733.4	35626.2

410000	39470.4	29466.2	23121.4	25117.2	36113.4
420000	39922.6	29757.8	23574.6	25460.6	36680
430000	40346.4	30393	23875	25775.4	37544.2
440000	40822.8	30808.8	24180.2	26096.4	38251.2
450000	41352.6	31110.4	24448.2	26410	39023.4
460000	42139.8	31750.4	24828.4	26795.4	39472.8
470000	42656.4	31923.2	25174.2	27094.4	40113.6
480000	43156.4	31989	25478	27416.8	40614
490000	43702.4		25752.8	27819	41029.6
500000	44170.2				41499

The above information is rendered below in graph form. It reveals that poetry has the overall largest number of types.

Figure 11: Prose, Grammar Chat-site, Plays and POEGRACHAPLA types



The graph reveals that from the 10,000 token mark to the 500,000 token point Poetry word types soar above all others. This may offer support to the high lexical density use in poetic language in general. The Poetry text type is followed consistently by POEGRACHAPLA until the end. From 130,000 up to 500,000 tokens Chat-site has the lowest number of types overall. Although Chat-site text has a mixture of Setswana and English words, typos, misspellings, and the general lack of standard spelling, the evidence shows that such language mixture does not translate into high word types.

POEGRACHAPLA has more types than Grammar, Plays and Chat-site texts but lower than Poetry text. The higher level of word types in POEGRACHAPLA suggests that a combination of text from a variety of text types in a corpus may result with

higher levels of types.

Next we measure: Prose, Newspaper, Hansard, Call-in etc (interviews and open radio programs) and PRONEWHANCAL. The results follow in Table 60.

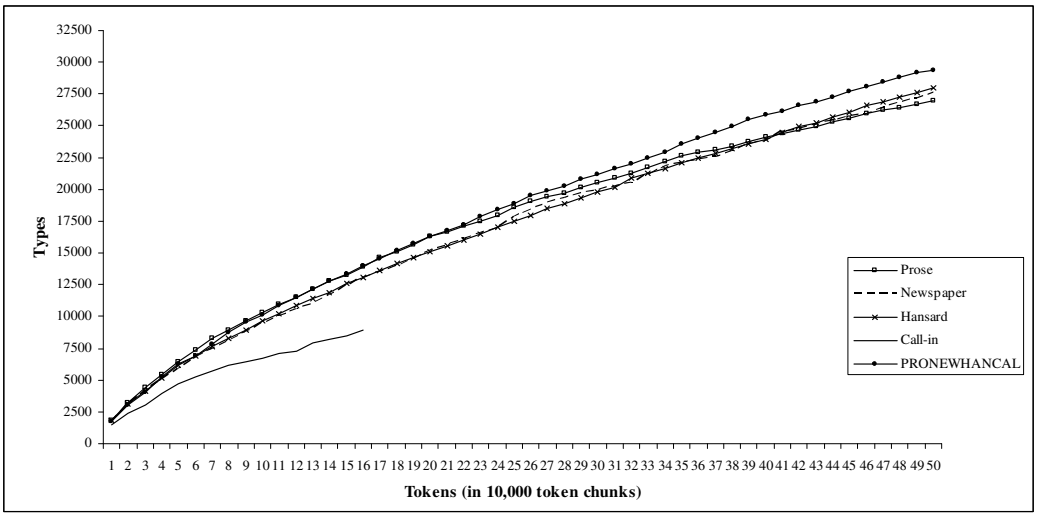
Table 60: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types

Tokens	Prose	Newspaper	Hansard	Call-in, etc	PRONEWHANCAL
10000	1852.6	1743.4	1880	1478.8	1794.6
20000	3216.4	3098.8	3133	2355.8	3076.4
30000	4427.4	4137.4	4145.2	3015.2	4091.8
40000	5430.8	5060.8	5198.8	3951.2	5274.2
50000	6488.8	5918	6178.4	4696	6254
60000	7348.8	6824.4	6929.6	5256.2	6882
70000	8242.8	7484.4	7644.2	5720.2	7781.2
80000	8926.4	8072.6	8328.4	6154.2	8764
90000	9694.8	8859.2	8942.2	6457.8	9543.8
100000	10283.8	9461.8	9644.2	6718.8	10151.8
110000	10939.2	10014	10210.2	7062	10881.8
120000	11477.6	10606.4	10854	7252.6	11501.4
130000	12123.2	11087.2	11404.6	7905.8	12156.4
140000	12783.6	11697.2	11918	8239	12793.4
150000	13242	12461.4	12577.4	8474.2	13378.8
160000	13931.2	13077.2	13108.2	8899	13988
170000	14608	13535.6	13631.6		14570.6
180000	15128.2	14015.6	14152.8		15159.4
190000	15634.6	14510.6	14619.6		15789.2
200000	16254.2	15182.8	15107.8		16270
210000	16686	15672	15587.2		16742.6
220000	17156	16113.2	16012		17257
230000	17465.4	16585.4	16480.4		17880
240000	17970	16981.2	17071		18379
250000	18557.8	17905.6	17462.6		18905.4
260000	19077.4	18372.8	17953.2		19490.8
270000	19408	18945.4	18464.8		19880.6
280000	19724.4	19319.4	18840		20229.4
290000	20181.4	19658	19293.4		20764.6
300000	20510	19960.8	19771.2		21213.4
310000	20919	20227.2	20191.8		21597
320000	21303.8	20540.6	20859.8		22044.2
330000	21731.2	21151	21243		22509.8
340000	22184.8	21780.6	21680		22913
350000	22630	22064.4	22062		23565.2
360000	22930	22340.4	22503.6		23985.8
370000	23138.2	22595.4	22840.2		24517
380000	23392.4	23056.4	23216.8		24925
390000	23795	23539.2	23609.8		25505.4
400000	24138.6	23809.2	23979.4		25884.2

410000	24441.4	24544.6	24518		26171.4
420000	24664.4	24801.8	24993		26568.8
430000	24947.8	25151.6	25266.4		26914
440000	25328	25452.6	25687.6		27252.6
450000	25603.6	25738.8	26073.2		27756.4
460000	25967	25947.8	26599.6		28080.2
470000	26201.8	26393.2	26915.6		28480
480000	26401.4	26853.8	27293.4		28863
490000	26716.2	27115.6	27622.6		29220.4
500000	26941	27607	28005		29367.4

Figure 12 renders the Table 60 results in a graphical form.

Figure 12: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types



From the beginning to the end, Call-in (interview and open radio program) display the lowest number of types compared to Prose, Newspaper and Hansard. This implies that individuals who call radio stations or are interviewed on radio and television, in general, use a limited vocabulary. Between 10,000 and 400,000 tokens Prose has the largest number of word types, after which Hansard word types lead until the 500,000 tokens point. The Hansard types display consistent increase up to the 500,000 tokens point where they are second to the PRONEWHANCAL types and close to the Newspaper types. This may be expected about Hansard text since Hansards document parliamentary debates which are on a variety of topics. The Newspaper types are the most unstable. At certain points they exceed the Hansard types and by the 500,000 tokens point they had exceeded the Prose types. Between 250,000 and 310,000 tokens

they exceed the Hansard types and between 410,000 and 500,000 tokens they exceed the Prose types. From 180,000 tokens PRONEWHANCAL types lead until 500,000 tokens. Since PRONEWHANCAL comprises texts from prose, newspaper, Hansard and call-in text, the high level of types that characterise it, may give support to the view that corpora compiled from a variety of text types have a higher lexical density.

We now turn to Science, Politics, Business, Religious and SCIPOLBUSREL texts. In the entire Setswana corpus, Business, Science and Politics have some of the smallest number of tokens. In terms of our 10,000 chunks they each have 100,000, 140,000 and 200,000 tokens respectively. Religious texts have 480,000 tokens. Below are the results of the calculation of word types for the five text types at comparable points.

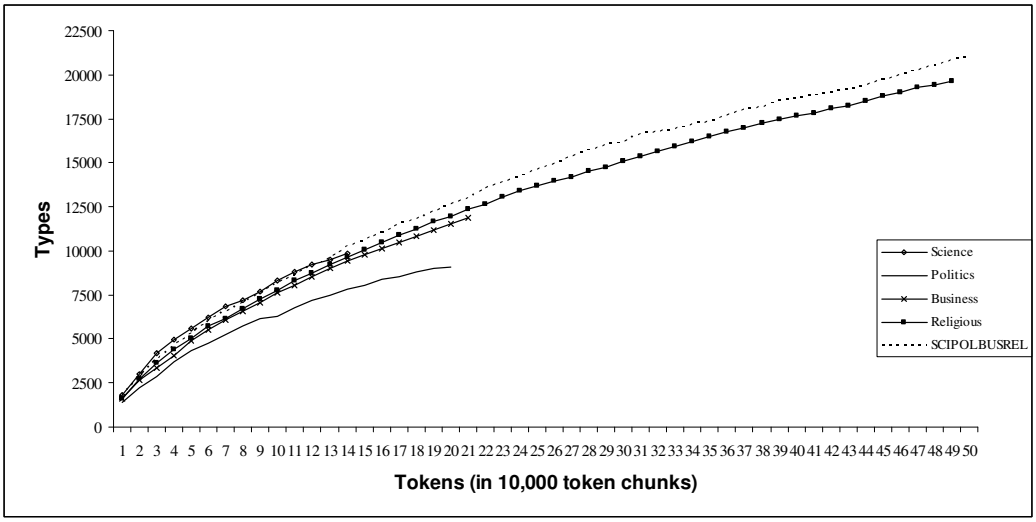
Table 61: Science, Politics, Business, Religious and SCIPOLBUSREL types

Tokens	Science	Politics	Business	Religious	SCIPOLBUSREL
10000	1806	1431	1629.2	1574.6	1751
20000	2972.8	2253	2647.4	2723	2864.6
30000	4167.8	2895	3364	3655.8	3811.4
40000	4970.4	3672	4067.6	4395.2	4614.6
50000	5615.4	4314	4874	5061.4	5309.6
60000	6218.2	4774.2	5507.8	5721.6	6004.2
70000	6841.8	5236.6	6046.6	6177.6	6582
80000	7219.6	5755.6	6587	6741.2	7055
90000	7714.4	6133.4	7070.6	7260	7607.2
100000	8281.8	6274.8	7619.6	7733.2	8172
110000	8778.6	6745.8		8326.4	8622.8
120000	9239.4	7213.8		8763.2	9184.8
130000	9517.2	7490		9229.4	9600
140000	9883	7814.4		9644.6	10262.4
150000		8038		10052.6	10643
160000		8357.8		10448	11008.2
170000		8534		10894.8	11498.6
180000		8780.6		11270.4	11809.4
190000		8997.2		11640.6	12253.4
200000		9063		11971.6	12674
210000				12378.2	12996.8
220000				12682.4	13583.6
230000				13051.6	13899.6
240000				13400.8	14259.4
250000				13672.2	14608.8
260000				13961.2	14973.2
270000				14197.8	15377.8
280000				14507.2	15740.6
290000				14771.4	15984.8
300000				15094.2	16210.6

310000				15399.2	16600.2
320000				15682.8	16767.2
330000				15918.6	16922
340000				16179.2	17171.4
350000				16487.6	17351.2
360000				16750.2	17665.6
370000				16988.6	18001.4
380000				17269.2	18152
390000				17499.6	18495.4
400000				17693.4	18658
410000				17833.4	18846.4
420000				18087.8	18976.8
430000				18264	19142.8
440000				18524	19446.2
450000				18818.8	19713.2
460000				18989.4	19968.2
470000				19259.6	20262.2
480000				19418.6	20529
490000				19646	20810.4
					21047

We plot the above information in the graph below.

Figure 13: Science, Politics, Business, Religious and SCIPOLBUSREL types



From 10,000 to 120,000 tokens Science text leads with the highest types after which it is overtaken by SCIPOLBUSREL which maintains the highest number of types until the 500,000 tokens mark. Not enough data however is available to track the development of the Science text up to the 500,000 tokens mark since it has only 140,000 tokens. From 150,000 Religious text type has the second largest number of types until at 500,000 tokens. Of all the text types Politics have the smallest number

of types. Since SCIPOLBUSREL leads with types between 130,000 and 500,000 token points, this may provide evidence that corpora compiled from a variety of text types do render higher levels of word types.

Having looked at Science, Politics, Business, Religious and SCIPOLBUSREL types we now look at the newspaper text type and measure its subcomponents.

7.2.1 Newspaper Components type/token

While we have looked at the genre of Newspaper text as a single unit above, we recognise that it has different components. This position is similar to that of Kovarik who argues that newspaper texts constitute a sublanguage – a version of a natural language which does not display all of the creativity of that natural language. "The newspaper sublanguage can be further constrained by subject matter to divide it into smaller, more manageable subsets" (Kovarik, 2000: 116/117).

The more manageable subsets that we have isolated in the Setswana newspapers are: Arts and Culture, Business, Letters, News and Sport. These are analysed in a similar manner as other components above. Similarly we give the components' types against token chunks at 10,000 token intervals and we subsequently plot these on a graph.

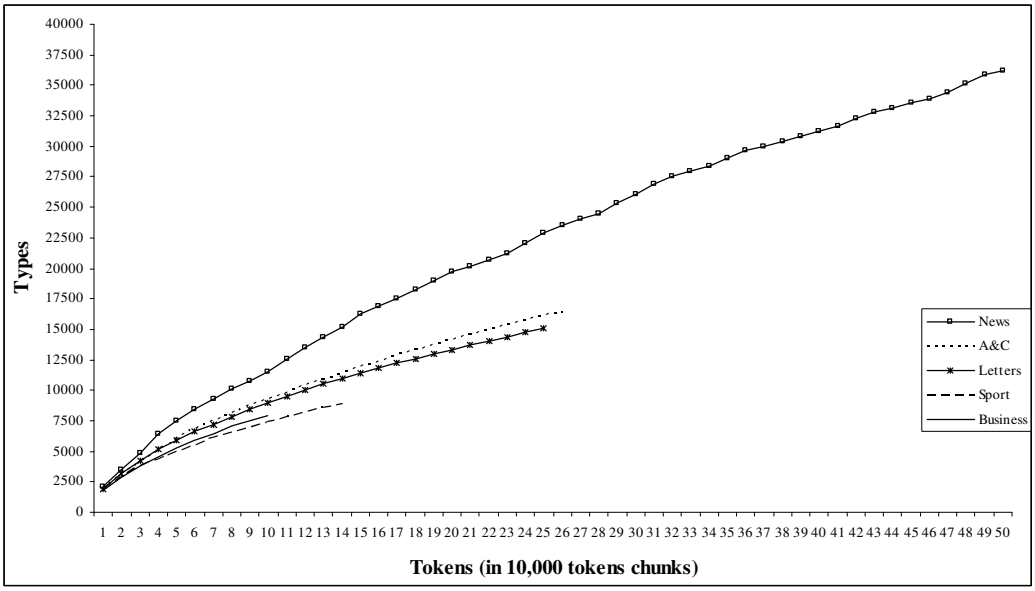
Table 62: Newspaper components types

Tokens	News	A&C	Letters	Sport	Business
10000	2159.2	1927	1901.4	1854.6	1812.2
20000	3530	3140.6	3195.4	2840	2875.8
30000	4886.4	4168.2	4248.4	3795.8	3820.4
40000	6422	5089.2	5151.2	4351.8	4561.4
50000	7516.8	6060.4	5939.6	4955.8	5269.6
60000	8435.8	6808.4	6678.4	5510.6	5922
70000	9239.2	7457.2	7221.2	6090	6459.2
80000	10112.2	8127	7855.6	6562	7028.6
90000	10761.4	8723	8460.4	6992.2	7511.4
100000	11519.4	9276	9021	7418	7937
110000	12554.6	9858.8	9541	7809.4	
120000	13525.2	10414.6	10011.4	8212.8	
130000	14325.2	10892.6	10513.4	8581.6	
140000	15167.4	11399.2	10931	8837.2	
150000	16272.6	11901.8	11350.2		
160000	16921.2	12348.8	11799.4		

170000	17569.8	12899.6	12213.2		
180000	18269.4	13333.2	12573.6		
190000	18944.6	13766.4	12984.8		
200000	19752.2	14134.6	13342.2		
210000	20121	14555.8	13682.2		
220000	20700.8	14989	14057		
230000	21171.4	15389.6	14390.4		
240000	22032.4	15760.6	14762.4		
250000	22940.6	16105	15111		
260000	23537	16392			
270000	24097				
280000	24456.8				
290000	25307				
300000	26055.8				
310000	26863.2				
320000	27503.2				
330000	28018				
340000	28352.6				
350000	28987.6				
360000	29641.2				
370000	29943.2				
380000	30432				
390000	30771.4				
400000	31280.8				
410000	31621.4				
420000	32337.2				
430000	32829.8				
440000	33145.4				
450000	33579.4				
460000	33846.6				
470000	34363.4				
480000	35105.8				
490000	35927.8				
500000	36206				

Table 62 data in graphical form is presented below in Figure 14.

Figure 14: Newspaper components types

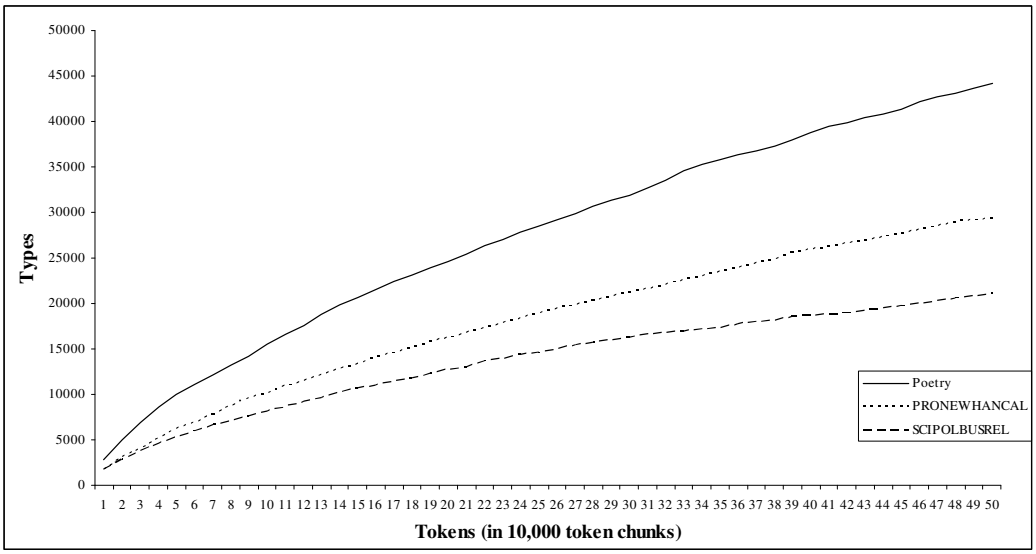


The graph clearly reveals that News types soar above the rest. This is probably because of the different kinds of subjects covered in news compared to Business, Arts and Culture, Letters and Sport. News report on a variety of subjects which we suggest would be responsible for the high number of types compared to the other sections of the newspaper. News word types significantly begin to break away at 20,000 tokens with 3,530 types. Arts and Culture followed by Letters follow News in the number of types, although Letters types are never far removed from the Arts and Culture ones. Sport has the lowest types consistently compared to other text types.

7.3 Conclusion of type-token measurements

In the above experiments we have measured types of various text types at 10,000 tokens intervals. We found that Poetry, PRONEWHANCAL and SCIPOLBUSREL have the largest overall types in general. When we compare these three we find that Poetry leads PRONEWHANCAL and SCIPOLBUSREL. This is reflected in Figure 15 below).

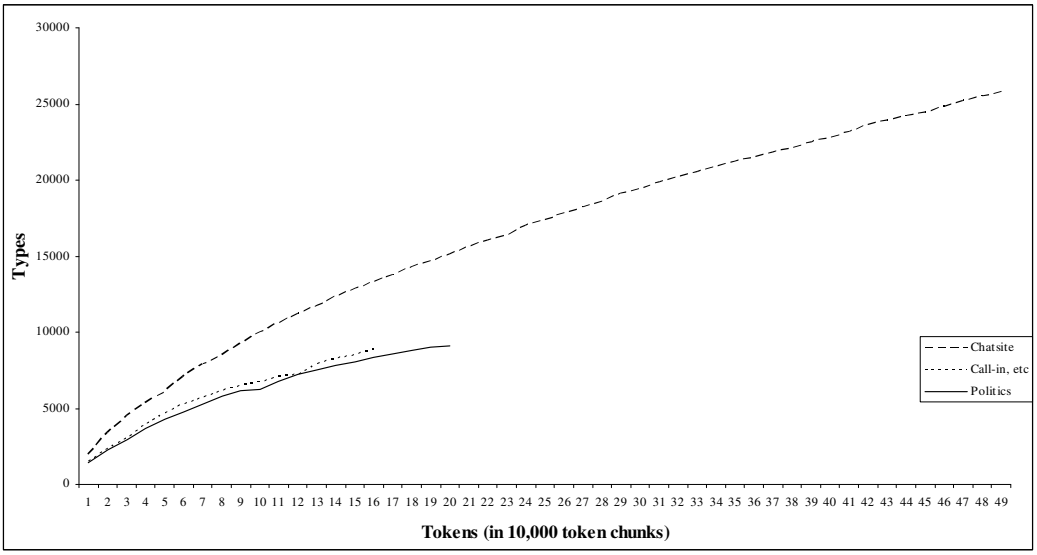
Figure 15: Comparison of the three overall top text types



Overall Poetry text has the largest number of types at most of the 10,000 tokens intervals followed by PRONEWHANCAL and SCIPOLBUSREL respectively. We conclude that poetry uses a wide vocabulary compared to other text types. Given the high number of types in PRONEWHANCAL and SCIPOLBUSREL, and in POEGRACHAPLA, we can safely conclude that combining texts from a variety of text types to compile a corpus leads to a higher number of types.

We have also seen that text types with the lowest types are Chat-site, Call-in and Politics.

Figure 16: Comparison of the three overall lowest text types



Politics have the lowest types overall, followed by Call-in and Chat-site. This suggests that these three use a limited vocabulary when compared with other text types. It should be emphasised, however, that while certain text types contribute the lowest number of types, such text types are not less important or less significant to corpus compilation, since text types with the lowest number of types also do contribute unique words which would enrich a headword list.

Next, we test how frequency lists from different text types and the frequency lists from the three compilations PRONEWHANCAL, SCIPOLBUSREL, and POEGRACHAPLA perform when juxtaposed to the frequency lists generated from the whole corpus.

7.4 A comparison of the top 100 tokens

Below the whole Setswana corpus' most frequent 100 words are used as a standard against which to compare Poetry, Grammar, Chat-site, Plays, Prose, Spoken, Miscellaneous, Science, Politics, Business, Religious, Newspaper, PRONEWHANCAL, SCIPOLBUSREL and POEGRACHAPLA's most frequent 100 words. The purpose of the experiment is to determine the differences between the top 100 words extracted from a mixture of text types (i.e. the whole corpus) and that of

individual corpus text types that form part of the entire corpus. We also wish to determine how the top 100 words of the whole corpus compare to a limited combination of text types as represented in PRONEWHANCAL, SCIPOLBUSREL and POEGRACHAPLA.

There are at least two approaches that could be adopted to extract the 100 frequent tokens from the entire corpus. Raw frequency counts could be ordered from the most frequent to the least frequent. Such an approach's results are in Table 63.

Table 63: Top 100 most frequent tokens in the whole corpus

N	Word	Freq.	Texts
1	a	676,657	2,845
2	go	413,587	2,793
3	e	403,383	2,806
4	le	354,572	2,772
5	o	327,853	2,788
6	ba	311,646	2,699
7	ka	287,741	2,749
8	ke	241,249	2,734
9	ya	225,776	2,752
10	mo	191,304	2,733
11	re	157,637	2,523
12	ga	148,640	2,667
13	fa	141,858	2,630
14	se	131,599	2,540
15	gore	124,504	2,639
16	di	122,905	2,610
17	ne	96,518	2,279
18	wa	94,050	2,613
19	tsa	91,423	2,571
20	sa	80,426	2,569
21	i	71,499	1,385
22	tse	68,069	2,451
23	kwa	67,921	2,432
24	bo	61,587	2,478
25	mme	59,585	2,401
26	tla	54,537	2,191
27	la	48,330	2,383
28	nna	42,931	2,280
29	yo	36,420	2,002
30	fela	36,309	2,203
31	gagwe	34,149	1,634
32	na	32,970	2,107
33	bona	32,779	1,777
34	bone	30,405	2,214
35	jwa	27,876	2,005
36	jaaka	27,285	2,063
37	batho	26,891	1,979
38	the	24,563	771
39	lo	24,264	951
40	itse	23,349	1,537
41	ntse	23,319	1,614
42	motho	21,357	1,572
43	teng	20,581	1,798
44	to	20,405	632
45	mongwe	20,121	1,681
46	neng	19,556	1,593
47	dira	19,365	1,816
48	jalo	18,207	1,868
49	ene	18,031	1,506
50	bua	17,075	1,420
51	tswa	16,847	1,772
52	rona	16,286	1,488
53	me	16,236	878
54	thata	15,630	1,731
55	kgotsa	15,522	1,231
56	pele	14,992	1,633
57	and	14,081	693
58	of	13,823	836
59	morago	13,687	1,528
60	posted	13,636	327
61	gago	13,515	835
62	kana	13,508	1,387
63	jaanong	13,278	1,323
64	eng	13,277	1,192
65	tshwanetse	12,812	1,310
66	bana	12,553	1,147
67	nako	12,270	1,465
68	batla	11,632	1,382
69	you	11,573	413
70	gape	11,563	1,577
71	yone	11,540	1,512

72	madi	11,393	1,290
73	nngwe	11,069	1,268
74	setse	10,905	1,221
75	ngwana	10,883	827
76	monna	10,832	774
77	tsaya	10,692	1,352
78	leng	10,656	1,238
79	bangwe	10,585	1,611
80	gone	10,531	1,417
81	bile	10,477	1,333
82	ntlha	10,323	1,093
83	dilo	10,248	1,296
84	jaana	10,176	1,247
85	wena	10,070	813
86	tsena	10,056	1,079

87	on	10,032	530
88	is	9,982	541
89	rile	9,982	906
90	utlwa	9,929	932
91	be	9,827	773
92	jang	9,816	1,314
93	tiro	9,770	1,170
94	kgosi	9,668	477
95	sengwe	9,623	1,146
96	tota	9,576	1,381
97	jo	9,532	1,299
98	lefatshe	9,438	1,335
99	botswana	9,433	1,502
100	sentle	9,418	1,252

The results of Table 63 are useful and may be used in the compilation of a headword list. The results are listed on the basis of frequency of occurrence in the entire Setswana corpus. *A* is the most frequent token in the corpus occurring 676,657 times and found in 2,845 texts. *Sentle* occupies the 100th word spot with 9,418 occurrences in 1,252 texts. However Leech et al. (2001: 17) contend that "simple word frequency counts can be misleading." This is because,

> If a word has a high frequency count, the user may infer, because the compilers have attempted to build a large, maximally representative corpus, that the word has a similarly high occurrence in the … language as a whole. However this may be a false inference. It is possible that the word has a high frequency not because it is widely used in the language as a whole but because it is 'overused' in a much smaller number of texts, or parts of texts, within the corpus (Leech et al., 2001: 17).

To address this matter they suggest dispersion statistics (Range (Ra) and Dispersion (Disp)) which show whether a word is widely spread because it occurs in many of the text samples or whether it is because of high usage in only a few samples. They argue that,

> Frequent words with high dispersion values may be considered to have high currency in the language as a whole; high frequencies associated with low dispersion values should, in contrast, be treated with caution (Leech et al.,

2001: 18).

We will not explore any further the complexities of Leech et al.'s statistics, but we discussed them since they bare close semblance to Scott's (2004-2006: 109) Simple Consistency Analysis (SCA). SCA calculates words which recur consistently in lots of texts of a given genre and orders them on the basis of their spread. What SCA does is therefore to calculate word spread. SCA results are given on the basis of the number of texts the words occur in. The results are given in the word-list, for instance Table 64, in a column headed "Texts" which shows the calculated number of texts each word occurred in (the maximum number being the total number of text-files used for the word-list).

SCA is dependent on the number of text-files. The words occurring in the largest number of text files are listed at the top, while the ones occurring in fewer texts occur lower in the list. In Table 64 the top 100 words of the Setswana corpus are given on the basis of SCA measurement.

Table 64: Top 100 words: Simple Consistency Analysis results

N	Word	Freq.	%	Texts
1	a	676,657	5.22	2,845
2	e	403,383	3.11	2,806
3	go	413,587	3.19	2,793
4	o	327,853	2.53	2,788
5	le	354,572	2.74	2,772
6	ya	225,776	1.74	2,752
7	ka	287,741	2.22	2,749
8	ke	241,249	1.86	2,734
9	mo	191,304	1.48	2,733
10	ba	311,646	2.40	2,699
11	ga	148,640	1.15	2,667
12	gore	124,504	0.96	2,639
13	fa	141,858	1.09	2,630
14	wa	94,050	0.73	2,613
15	di	122,905	0.95	2,610
16	tsa	91,423	0.71	2,571
17	sa	80,426	0.62	2,569
18	se	131,599	1.02	2,540
19	re	157,637	1.22	2,523
20	bo	61,587	0.48	2,478
21	tse	68,069	0.53	2,451
22	kwa	67,921	0.52	2,432
23	mme	59,585	0.46	2,401
24	la	48,330	0.37	2,383
25	nna	42,931	0.33	2,280
26	ne	96,518	0.74	2,279
27	bone	30,405	0.23	2,214
28	fela	36,309	0.28	2,203
29	tla	54,537	0.42	2,191
30	na	32,970	0.25	2,107
31	jaaka	27,285	0.21	2,063
32	jwa	27,876	0.22	2,005
33	yo	36,420	0.28	2,002
34	batho	26,891	0.21	1,979
35	jalo	18,207	0.14	1,868
36	dira	19,365	0.15	1,816
37	teng	20,581	0.16	1,798
38	bona	32,779	0.25	1,777
39	tswa	16,847	0.13	1,772
40	thata	15,630	0.12	1,731
41	mongwe	20,121	0.16	1,681
42	gagwe	34,149	0.26	1,634
43	pele	14,992	0.12	1,633
44	ntse	23,319	0.18	1,614
45	bangwe	10,585	0.08	1,611
46	neng	19,556	0.15	1,593
47	gape	11,563	0.09	1,577
48	motho	21,357	0.16	1,572
49	itse	23,349	0.18	1,537
50	morago	13,687	0.11	1,528
51	yone	11,540	0.09	1,512

52	ene	18,031	0.14	1,506
53	botswana	9,433	0.07	1,502
54	rona	16,286	0.13	1,488
55	nako	12,270	0.09	1,465
56	bua	17,075	0.13	1,420
57	gone	10,531	0.08	1,417
58	kana	13,508	0.10	1,387
59	i	71,499	0.55	1,385
60	batla	11,632	0.09	1,382
61	tota	9,576	0.07	1,381
62	tsaya	10,692	0.08	1,352
63	one	9,057	0.07	1,343
64	lefatshe	9,438	0.07	1,335
65	bile	10,477	0.08	1,333
66	jaanong	13,278	0.10	1,323
67	jang	9,816	0.08	1,314
68	tshwanetse	12,812	0.10	1,310
69	jo	9,532	0.07	1,299
70	dilo	10,248	0.08	1,296
71	madi	11,393	0.09	1,290
72	nngwe	11,069	0.09	1,268
73	sentle	9,418	0.07	1,252
74	dingwe	8,177	0.06	1,251
75	jaana	10,176	0.08	1,247
76	sena	7,506	0.06	1,243

77	leng	10,656	0.08	1,238
78	kgotsa	15,522	0.12	1,231
79	setse	10,905	0.08	1,221
80	rre	8,383	0.06	1,212
81	nne	8,023	0.06	1,204
82	eng	13,277	0.10	1,192
83	tiro	9,770	0.08	1,170
84	batswana	5,902	0.05	1,167
85	bana	12,553	0.10	1,147
86	sengwe	9,623	0.07	1,146
87	supa	6,033	0.05	1,145
88	gaborone	3,326	0.03	1,128
89	eo	7,017	0.05	1,125
90	ngwaga	5,216	0.04	1,116
91	dumela	4,380	0.03	1,096
92	ntlha	10,323	0.08	1,093
93	tsone	6,274	0.05	1,093
94	tsotlhe	7,330	0.06	1,086
95	tsena	10,056	0.08	1,079
96	gongwe	8,412	0.06	1,053
97	mokgosi	2,704	0.02	1,046
98	sepe	7,045	0.05	1,040
99	seka	3,713	0.03	1,033
100	raya	9,132	0.07	1,018

Table 64 in the first column shows the rank of a word, followed by the word ranked, which is followed by the word's frequency in the whole corpus. The fourth column is of the word's frequency as a percentage of the corpus, followed by the SCA score which is the number of texts each word appears in.

When we compare the SCA results with those of raw frequencies, we find that all the English words that appear in the top 100 raw frequencies words no longer appear in the 100 SCA results. The English words are, *I* (21), *on* (87), *you* (69), *posted* (60), *is* (88), *be* (91), *the* (39), *of* (58). English words are not spread throughout the corpus but are limited to a few files, that is why they do not appear amongst the 100 SCA results.

We will however use raw frequency measure as a standard against which we measure the most frequent words from different text types to make our study comparable to many studies in the field which use raw frequencies and not dispersion results. Additionally, we use raw frequency counts so that later in this chapter we could compare our lists and results with the BNC lists. In practical dictionary compilation, Leech et al.'s guidance that both raw frequencies and dispersion results should be considered in the selection of headwords.

Our study is similar to that of Sharoff (2006) in which he investigates the possibility to develop a BNC-like corpus for a number of different languages (Chinese, English, German, Romanian, Ukrainian and Russian). He also evaluates the collected corpora using the composition of resulted corpora and their frequency lists for some of the languages (English, German and Russian). He compares the internet compiled corpus with large available balanced English and Russian corpora. For English he used the BNC, for Russian, he used the Russian Reference Corpus (RRC). It is particularly the sections on corpus comparison by Sharoff that interest us.

We compare each of the 15 text types' most frequent 100 types against those of the whole corpus' most frequent 100 tokens. The 15 groups are the following:

A	B	C
Poetry	Science	Prose
Grammar	Politics	Newspaper
Chat-site	Business	Hansard
Plays	Religious	Call-in
POEGRACHAPLA	SCIPOLBUSREL	PRONEWHANCAL

Below we give results in Tables 65, 66, and 67 and follow the results with a discussion. Where a word in the top 100 words of the whole corpus is not in the top 100 words of a text type we indicate such absence by "–" followed by a bracketed number to show the rank it occupies in the list.

Table 65: Poetry, Grammar, Chat-site, Plays and POEGRACHAPLA

Whole corpus	Poetry	Grammar	Chat-site	Plays	POEGRACHAPLA
1. a	1	1	4	1	1
2. e	5	3	31	7	5
3. go	4	2	18	4	2
4. o	6	6	14	3	4
5. le	3	4	20	6	6
6. ya	9	7	58	13	10
7. ka	7	5	41	5	7
8. mo	16	9	60	11	12
9. ke	2	11	19	2	3
10. ba	8	8	26	8	8
11. ga	11	15	51	10	9
12. fa	17	13	100	14	15
13. gore	33	18	43	15	19

14. di	12	10	87	18	14
15. wa	14	19	92	16	16
16. tsa	18	17	- (153)	25	22
17. sa	15	20	- (110)	20	17
18. se	13	12	- (123)	12	13
19. re	10	16	34	9	11
20. tse	32	14	- (275)	29	30
21. bo	19	24	74	21	23
22. kwa	24	23	- (327)	22	29
23. mme	25	22	- (216)	24	33
24. la	20	27	- (240)	34	38
25. nna	23	28	93	23	27
26. ne	27	21	- (192)	17	26
27. bone	37	51	- (202)	45	50
28. tla	21	25	- (162)	19	25
29. fela	26	34	- (189)	30	42
30. jaaka	25	30	- (407)	38	39
31. na	31	29	- (233)	33	41
32. jwa	34	32	- (686)	60	55
33. batho	35	38	- (176)	43	48
34. jalo	- (121)	52	- (499)	65	73
35. yo	38	26	- (221)	27	35
36. teng	52	50	- (277)	49	63
37. tswa	45	58	- (460)	54	70
38. dira	81	35	- (456)	57	74
39. thata	58	36	- (591)	61	59
40. bona	28	33	- (203)	28	46
41. mongwe	86	43	- (365)	62	86
42. pele	62	75	- (751)	79	93
43. gagwe	53	45	- (13,519)	36	51
44. bangwe	72	88	- (428)	- (114)	- (128)
45. gape	- (134)	68	- (342)	- (111)	- (109)
46. ntse	54	57	- (402)	37	57
47. neng	- (261)	78	- (1,201)	44	- (108)
48. botswana	- (50)	- (184)	71	- (1,423)	84
49. yone	82	55	- (455)	- (139)	83
50. motho	30	40	- (201)	35	53
51. morago	76	72	- (992)	82	- (106)
52. itse	36	53	- (205)	31	49
53. rona	46	76	- (182)	41	66
54. nako	- (106)	62	- (512)	- (103)	- (104)
55. ene	64	- (105)	- (366)	48	80
56. gone	- (173)	- (108)	- (259)	85	- (124)
57. bua	56	67	- (251)	42	61
58. kana	63	81	- (184)	50	79
59. batla	67	91	- (253)	51	78
60. lefatshe	43	- (202)	- (1,311)	- (131)	- (183)
61. tota	77	- (183)	- (226)	69	- (178)
62. tsaya	- (110)	93	- (695)	89	- (132)
63. madi	- (107)	- (130)	- (423)	87	- (179)
64. one	- (143)	- (112)	53	- (127)	76
65. tshwanetse	- (440)	84	- (782)	83	- (206)
66. bile	- (155)	- (129)	- (726)	72	- (133)

67. jaanong	- (139)	- (121)	- (351)	52	98
68. jang	- (148)	69	- (485)	74	- (105)
69. dilo	- (104)	82	- (445)	91	- (134)
70. nngwe	- (168)	56	- (1,363)	- (151)	- (190)
71. dingwe	- (183)	47	- (1,383)	- (207)	- (176)
72. rre	- (137)	- (212)	- (901)	- (105)	- (205)
73. sentle	- (171)	86	- (558)	70	- (120)
74. jaana	- (321)	- (128)	- (385)	64	100
75. kgotsa	83	31	- (963)	- (130)	- (157)
76. sena	- (329)	- (199)	- (571)	- (172)	- (250)
77. setse	- (103)	- (142)	- (724)	73	- (141)
78. nne	94	99	- (795)	- (126)	- (172)
79. tiro	- (114)	65	- (1,581)	81	- (195)
80. leng	- (150)	80	- (730)	- (112)	- (177)
81. batswana	- (105)	- (384)	86	- (562)	- (142)
82. supa	- (178)	70	- (3,772)	- (206)	- (295)
83. jo	- (215)	- (101)	- (1,108)	- (101)	- (186)
84. gaborone	- (181)	- (651)	- (230)	- (623)	- (228)
85. sengwe	- (142)	49	- (694)	71	- (151)
86. ngwaga	- (191)	- (198)	- (3,467)	- (358)	- (334)
87. eo	- (239)	- (203)	- (931)	- (166)	- (278)
88. bana	41	64	- (279)	80	82
89. tsone	- (129)	- (107)	- (868)	189	- (146)
90. eng	- (144)	42	- (295)	39	52
91. mokgosi	- (304)	- (2,022)	- (28,185)	- (773)	- (772)
92. tsotlhe	73	- (127)	- (2,762)	- (152)	- (182)
93. ntlha	- (113)	60	- (1,996)	- (106)	- (137)
94. dumela	43	- (330)	- (237)	- (117)	- (192)
95. tlhalosa	- (555)	79	- (6,527)	- (520)	- (361)
96. tsena	68	- (191)	- (556)	59	91
97. gongwe	- (125)	74	- (529)	100	- (184)
98. mangwe	- (302)	- (134)	- (2,469)	- (428)	- (309)
99. gompieno	- (257)	- (287)	- (1,006)	- (113)	- (212)
100.seka	- (429)	- (943)	- (491)	- (398)	- (576)
Total	**62**	**74**	**21**	**72**	**59**

Table 66: Science, Politics, Business, Religious and SCIPOLBUSREL

Whole corpus	Science	Politics	Business	Religious	SCIPOLBUSREL
1. a	5	5	1	1	4
2. e	4	4	2	5	3
3. go	1	1	3	4	1
4. o	11	11	5	6	9
5. le	2	2	4	2	2
6. ya	3	3	6	8	5
7. ka	6	6	8	7	7
8. mo	9	9	11	9	8
9. ke	14	14	15	10	12
10. ba	7	7	7	3	6
11. ga	13	13	14	12	13
12. fa	17	17	13	14	17
13. gore	15	19	16	29	16

14. di	10	10	10	16	11
15. wa	19	15	18	13	18
16. tsa	8	8	9	15	10
17. sa	18	22	22	21	22
18. se	16	16	17	17	15
19. re	35	38	21	11	20
20. tse	12	12	12	25	14
21. bo	25	27	27	35	25
22. kwa	20	28	20	19	21
23. mme	23	33	23	20	24
24. la	21	21	19	23	23
25. nna	26	26	32	32	27
26. ne	45	72	26	27	28
27. bone	42	60	28	36	30
28. tla	28	20	24	18	19
29. fela	46	47	33	49	39
30. jaaka	41	43	37	41	36
31. na	33	31	34	38	29
32. jwa	22	39	25	34	26
33. batho	24	36	39	50	33
34. jalo	52	67	40	54	44
35. yo	82	52	65	28	38
36. teng	37	66	42	42	45
37. tswa	53	78	53	78	50
38. dira	36	29	31	48	31
39. thata	63	98	52	60	62
40. bona	39	55	56	46	42
41. mongwe	64	24	55	44	32
42. pele	69	45	67	40	58
43. gagwe	- (290)	- (122)	91	26	41
44. bangwe	- (125)	- (147)	87	- (179)	- (121)
45. gape	64	93	57	100	64
46. ntse	- (102)	- (193)	80	75	85
47. neng	- (188)	- (136)	68	58	80
48. botswana	27	92	30	- (381)	47
49. yone	61	- (170)	44	- (250)	76
50. motho	77	63	- (112)	76	71
51. morago	- (101)	80	73	- (104)	90
52. itse	- (139)	- (351)	- (236)	88	- (145)
53. rona	- (128)	- (127)	64	43	46
54. nako	68	82	61	- (206)	69
55. ene	84	- (624)	79	53	84
56. gone	259	- (497)	- (143)	- (194)	- (249)
57. bua	- (345)	- (300)	- (311)	67	89
58. kana	- (498)	- (152)	- (102)	- (116)	- (130)
59. batla	- (143)	- (284)	- (152)	- (178)	- (134)
60. lefatshe	48	- (120)	50	57	48
61. tota	- (449)	- (434)	- (194)	- (417)	- (230)
62. tsaya	97	- (101)	94	- (113)	91
63. madi	80	91	29	- (107)	57
64. one	- (137)	- (290)	- (116)	- (265)	- (177)
65. tshwanetse	32	25	38	- (162)	35
66. bile	- (354)	85	- (101)	- (190)	- (146)

67. jaanong	- (327)	- (288)	- (186)	86	- (155)
68. jang	92	- (210)	- (163)	- (117)	- (112)
69. dilo	98	- (134)	- (134)	93	- (106)
70. nngwe	75	30	62	- (109)	54
71. dingwe	54	53	63	- (279)	66
72. rre	- (534)	- (2,527)	74	- (348)	- (227)
73. sentle	- (174)	- (196)	- (136)	- (270)	- (141)
74. jaana	- (634)	- (262)	- (214)	- (115)	- (135)
75. kgotsa	30	18	45	- (322)	34
76. sena	- (209)	- (370)	- (160)	95	- (158)
77. setse	- (332)	- (536)	- (161)	- (373)	- (243)
78. nne	90	- (114)	92	63	94
79. tiro	73	54	41	- (142)	63
80. leng	95	77	93	89	79
81. batswana	- (181)	- (738)	- (148)	- (1,682)	- (334)
82. supa	- (122)	- (232)	99	- (184)	- (125)
83. jo	74	89	60	74	51
84. gaborone	- (471)	- (1,110)	- (169)	- (10,298)	- (514)
85. sengwe	- (262)	- (112)	- (219)	- (133)	- (133)
86. ngwaga	- (121)	- (154)	46	- (188)	88
87. eo	- (182)	100	- (142)	- (221)	- (175)
88. bana	- (176)	- (335)	- (253)	- (123)	- (182)
89. tsone	94	- (213)	83	- (374)	- (157)
90. eng	- (237)	- (274)	- (277)	- (114)	- (140)
91. mokgosi	- (2,945)	- (3,850)	- (135)	- (844)	- (707)
92. tsotlhe	(113)	99	98	51	70
93. ntlha	62	94	- (123)	64	68
94. dumela	- (700)	- (339)	- (301)	- (543)	- (405)
95. tlhalosa	- (316)	- (549)	77	- (1,057)	- (162)
96. tsena	- (382)	- (201)	- (403)	- (138)	- (293)
97. gongwe	- (273)	- (329)	- (231)	81	- (215)
98. mangwe	- (123)	- (137)	82	- (443)	- (129)
99. gompieno	- (547)	- (612)	- (235)	- (425)	- (303)
100.seka	- (112)	- (303)	- (187)	- (668)	- (170)
Total	**64**	**59**	**71**	**61**	**68**

Table 67: Prose, Hansard, Call-in, Newspaper and PRONEWHANCAL

Whole corpus	Prose	Hansard	Call-in	Newspaper	PRONEWHANCAL
1. a	1	1	1	1	1
2. e	4	2	7	2	3
3. go	2	6	4	4	2
4. o	5	- (163)	9	3	10
5. le	3	- (108)	3	6	7
6. ya	10	13	12	8	12
7. ka	7	9	10	7	8
8. mo	9	11	11	10	11
9. ke	8	7	2	9	5
10. ba	6	3	5	5	4
11. ga	12	15	14	14	14
12. fa	11	10	13	12	13
13. gore	16	4	8	13	9

14. di	17	12	17	16	16
15. wa	18	42	20	18	21
16. tsa	21	20	30	17	23
17. sa	19	25	26	21	20
18. se	15	14	15	15	15
19. re	14	5	6	11	6
20. tse	25	18	21	22	22
21. bo	23	16	18	23	18
22. kwa	20	26	35	20	24
23. mme	22	17	19	24	19
24. la	29	44	34	25	29
25. nna	28	35	27	31	27
26. ne	13	21	24	19	17
27. bone	40	24	29	26	28
28. tla	24	39	23	28	25
29. fela	31	22	25	27	26
30. jaaka	37	52	54	35	39
31. na	33	37	28	34	30
32. jwa	41	- (154)	87	30	48
33. batho	43	29	44	32	32
34. jalo	48	69	59	36	45
35. yo	30	55	46	33	31
36. teng	46	38	36	41	40
37. tswa	45	93	63	52	66
38. dira	58	46	55	39	49
39. thata	55	- (104)	71	46	61
40. bona	27	61	43	47	37
41. mongwe	44	- (116)	58	40	57
42. pele	52	- (121)	88	56	70
43. gagwe	26	- (153)	- (104)	37	47
44. bangwe	97	87	82	48	84
45. gape	63	- (118)	75	63	75
46. ntse	32	45	39	53	33
47. neng	36	92	- (103)	42	51
48. botswana	- (799)	75	91	38	64
49. yone	- (103)	41	62	49	52
50. motho	38	- (110)	76	50	67
51. morago	47	- (199)	- (114)	57	88
52. itse	34	54	49	58	42
53. rona	83	33	33	45	36
54. nako	59	- (128)	72	61	77
55. ene	42	- (111)	64	43	72
56. gone	86	48	48	71	56
57. bua	49	32	45	72	34
58. kana	74	40	41	59	44
59. batla	61	100	61	88	86
60. lefatshe	- (185)	78	- (109)	51	90
61. tota	92	79	53	74	69
62. tsaya	73	74	69	81	68
63. madi	- (113)	57	70	44	60
64. one	- (140)	63	68	86	87
65. tshwanetse	100	73	81	64	76
66. bile	67	- (132)	90	69	83

67. jaanong	72	23	31	78	35
68. jang	82	- (115)	52	93	74
69. dilo	90	62	67	80	81
70. nngwe	93	- (134)	- (171)	85	- (120)
71. dingwe	- (131)	- (131)	- (139)	84	- (128)
72. rre	- (107)	- (215)	37	55	58
73. sentle	77	- (122)	79	94	82
74. jaana	84	53	57	95	55
75. kgotsa	70	- (551)	- (336)	65	- (123)
76. sena	- (108)	- (124)	- (127)	79	- (113)
77. setse	50	- (123)	- (111)	90	80
78. nne	- (137)	84	93	- (106)	99
79. tiro	- (114)	- (143)	- (138)	73	- (108)
80. leng	- (117)	28	32	- (111)	46
81. batswana	- (712)	58	77	67	62
82. supa	- (190)	- (213)	- (243)	89	- (151)
83. jo	- (109)	- (137)	- (121)	76	- (109)
84. gaborone	- (437)	- (332)	- (107)	- (101)	- (170)
85. sengwe	87	98	96	100	100
86. ngwaga	- (281)	- (332)	- (211)	70	- (143)
87. eo	- (126)	- (152)	- (113)	83	- (115)
88. bana	64	83	47	62	78
89. tsone	- (182)	68	- (108)	92	- (103)
90. eng	54	91	56	- (114)	59
91. mokgosi	- (916)	- (11,066)	- (1,346)	68	- (354)
92. tsotlhe	- (135)	- (252)	- (200)	- (144)	- (147)
93. ntlha	69	- (162)	- (178)	- (105)	- (111)
94. dumela	- (186)	- (305)	- (194)	- (186)	- (221)
95. tlhalosa	- (244)	- (394)	- (350)	96	- (204)
96. tsena	62	- (105)	97	- (126)	- (101)
97. gongwe	- (122)	64	74	- (102)	94
98. mangwe	- (290)	- (180)	- (203)	- (109)	- (191)
99. gompieno	- (192)	81	83	- (118)	- (102)
100. seka	- (525)	- (129)	86	- (103)	- (106)
Total	**74**	**65**	**78**	**88**	**80**

Below we summarise the results of Tables 65, 66, and 67 which show the similarities between the different text types and the whole corpus.

A

Poetry	62
Grammar	74
Chat-site	21
Plays	72
POEGRACHAPLA	59

B

Science	64
Politics	59
Business	71
Religious	61
SCIPOLBUSREL	68

C

Prose	74
Hansard	65
Call-in	78
Newspaper	88
PRONEWHANCAL	80

Tables 65, 66, and 67 reveal that the results of POEGRACHAPLA, SCIPOLBUSREL and PRONEWHANCAL depend on the text types that constitute them. Since these

three are made from samples taken from other text types they largely reflect the general trend found in such text types. The results are summarised in tables A, B and C above. Let us illustrate this phenomenon by looking at POEGRACHAPLA which comprises texts from Poetry, Grammar, Chat-site and Plays whose top 100 token-similarity with the top 100 tokens of the whole Setswana corpus are 62, 74, 21 and 72 respectively. POEGRACHAPLA has a token similarity of 59 with the whole corpus. Texts that make up POEGRACHAPLA in general have smaller similarity with the top 100 texts of the entire Setswana corpus. Consequently POEGRACHAPLA has little similarity with the top 100 texts of the entire Setswana corpus. These results compare well with those of group C which are on average higher. Prose, Hansard, Call-in and Newspaper's top 100 token-similarities with the top 100 tokens of the whole Setswana corpus are 74, 65, 78 and 88 respectively. Consequently PRONEWHANCAL has a higher corpus similarity of 80 with the most frequent 100 tokens of the whole Setswana corpus.

The averages for the four text types in A, B and C are 57, 64 and 76 respectively.

We therefore conclude that it is not enough to have a corpus with a variety of text types to generate large numbers of types. It is also crucial that the individual text types that comprise a corpus should individually have higher levels of types, as in Group C.

7.4.1 Comparison of the top 100 tokens of spoken and written Setswana

We conclude this section of experiments by comparing the most frequent 100 words of the spoken and the written part of the corpus to that of the most frequent 100 words of the entire corpus. The results follow in Table 68.

Table 68: Comparison of written and spoken components to the whole corpus

Whole corpus		**Written language**	**Spoken language**
1.	a	1	1
2.	e	3	2
3.	go	2	4
4.	o	5	17
5.	le	4	16
6.	ya	8	12
7.	ka	7	8
8.	mo	10	11
9.	ke	9	6
10.	ba	6	3

11.	ga	11	13
12.	fa	12	10
13.	gore	16	7
14.	di	15	14
15.	wa	18	29
16.	tsa	19	13
17.	sa	20	25
18.	se	14	15
19.	re	13	5
20.	tse	23	20
21.	bo	24	18
22.	kwa	22	28
23.	mme	25	19
24.	la	27	42
25.	nna	28	31
26.	ne	17	23
27.	bone	35	26
28.	tla	26	32
29.	fela	31	24
30.	jaaka	36	58
31.	na	33	33
32.	jwa	34	- (136)
33.	batho	37	35
34.	jalo	48	62
35.	yo	29	49
36.	teng	45	38
37.	tswa	50	82
38.	dira	47	53
39.	thata	53	99
40.	bona	32	55
41.	mongwe	43	95
42.	pele	54	- (113)
43.	gagwe	30	- (133)
44.	bangwe	77	91
45.	gape	67	- (111)
46.	ntse	40	47
47.	neng	44	98
48.	botswana	99	81
49.	yone	82	45
50.	motho	41	- (103)
51.	morago	58	- (173)
52.	itse	39	52
53.	rona	57	36
54.	nako	63	- (110)
55.	ene	49	89
56.	gone	95	50
57.	bua	55	37
58.	kana	66	44
59.	batla	69	90
60.	lefatshe	96	85
61.	tota	- (102)	70
62.	tsaya	85	67
63.	madi	73	63
64.	one	- (109)	65
65.	tshwanetse	62	75
66.	bile	76	- (126)
67.	jaanong	78	27
68.	jang	90	93
69.	dilo	89	66
70.	nngwe	70	- (144)
71.	dingwe	- (111)	- (137)
72.	rre	- (108)	- (108)
73.	sentle	93	- (114)
74.	jaana	98	56
75.	kgotsa	52	- (394)
76.	sena	- (125)	- (134)
77.	setse	74	- (131)
78.	nne	- (121)	88
79.	tiro	86	- (143)
80.	leng	- (115)	34
81.	Batswana	- (179)	64
82.	supa	- (143)	- (190)
83.	jo	88	- (147)
84.	Gaborone	- (253)	- (232)
85.	sengwe	92	97
86.	ngwaga	- (159)	- (278)
87.	eo	- (134)	- (132)
88.	bana	64	72
89.	tsone	- (158)	74
90.	eng	61	71
91.	mokgosi	- (281)	- (3,863)
92.	tsotlhe	- (120)	- (221)
93.	ntlha	75	- (174)
94.	dumela	- (192)	- (262)
95.	tlhalosa	- (185)	- (342)
96.	tsena	87	- (106)
97.	gongwe	- (119)	68
98.	mangwe	- (213)	- (185)
99.	gompieno	- (191)	83
100.	seka	- (268)	- (125)
Total		**81**	**71**

Ninety four percent of the entire Setswana corpus is written language and only 6% is spoken language component. The effects of this phenomenon are reflected in the results. Eighty one of the top 100 words of the written component of the corpus are

found in the most frequent 100 words of the whole corpus. On the other hand, 71 words of the top 100 words of the spoken component are found amongst the most frequent 100 words in the entire corpus. The written component of the corpus is much more diverse in terms of the kind of texts it comprises while comparatively the spoken component is limited. This may explain the differences between the two.

7.4.2 Comparison of the top 100 tokens of spoken and written parts of the BNC

Below we compare our results with those of the BNC to determine the quality of our results in comparison to those of a larger balanced English corpus. The BNC lists in Tables 69-72 are from Kilgarriff's website (www.kilgarriff.co.uk). We start off by first listing the top 100 words of the whole BNC, and those of the written and spoken components. Table 69 lists the most frequent 100 words of the BNC. Table 70 comprises the most frequent 100 words of the written subcorpus. Table 71 contains the most frequent words of the context governed section of the spoken subcorpus. Table 72 gives the BNC's top 100 words of the demographic section of the spoken corpus (See Appendix 7 for the BNC POS codes).

Table 69: The BNC top 100 words of the whole corpus

Freq	**word**	**POS**
6187267	the	at0
2941444	of	Prf
2682863	and	Cjc
2126369	a	at0
1812609	in	prp
1620850	to	to0
1089186	it	pnp
998389	is	vbz
923948	was	vbd
917579	to	prp
884599	i	pnp
833360	for	prp
695498	you	Pnp
681255	he	Pnp
662516	be	Vbi
652027	with	Prp
647344	on	Prp
628999	that	Cjt
507317	by	Prp
478162	at	Prp
470943	are	Vbb
462486	not	xx0
461945	this	dt0
454096	but	Cjc
442545	's	Pos
433441	they	Pnp
426896	his	Dps
413532	from	Prp
409012	had	Vhd
380257	she	Pnp
372031	which	Dtq
370808	or	Cjc
358039	we	Pnp
343063	an	at0
332839	n't	xx0
325048	's	Vbz
322824	were	Vbd
286913	that	dt0
268723	been	Vbn
268490	have	Vhb
260919	their	Dps
259431	has	Vhz
255188	would	Vm0
249466	what	dtq
244822	will	vm0
239460	there	ex0
237089	if	cjs
234386	can	vm0
227737	all	dt0
218258	her	dps
208623	as	cjs
205432	who	pnq
205195	have	vhi
196635	do	vdb
194800	that	Cjt-dt0
190499	one	Crd
185277	said	Vvd
173414	them	Pnp

171174	some	dt0
168387	could	vm0
165014	him	pnp
163469	into	prp
163081	its	dps
160652	then	av0
156111	two	crd
155417	when	avq-cjs
154288	up	avp
153679	time	nn1
152619	my	dps
150958	out	avp
147324	so	av0
143405	did	vdd

142118	about	prp
138334	your	dps
137801	now	av0
137151	me	pnp
137026	no	at0
134029	more	av0
129451	other	aj0
125465	just	av0
125442	these	dt0
124884	also	av0
123916	people	nn0
123655	any	dt0
118699	first	ord
115994	only	av0
114655	new	aj0

113024	may	vm0
111538	very	av0
111236	should	vm0
111083	as	cjs-prp
108988	like	prp
108710	her	pnp
108618	than	cjs
106427	as	prp
101508	how	avq
96080	well	av0
95313	way	nn1
95001	our	dps
91583	as	av0

Table 70: The BNC top 100 words of the written corpus component

Freq	**word**	**POS**
5776384	the	at0
2789403	of	prf
2421302	and	cjc
1939617	a	at0
1695860	in	prp
1468146	to	to0
892937	is	vbz
845350	to	prp
839964	was	vbd
834957	it	pnp
768898	for	prp
606027	with	prp
605749	he	pnp
603178	be	vbi
590305	on	prp
580267	i	pnp
561041	that	cjt
490673	by	prp
435574	at	prp
426207	you	pnp
425898	's	pos
422562	are	vbb
413798	not	xx0
413737	his	dps
404140	this	dt0
390876	from	prp
389108	but	cjc
386510	had	vhd
349120	which	dtq
337345	she	pnp
336599	they	pnp
335976	or	cjc

323963	an	at0
294301	were	vbd
249636	we	pnp
247596	their	dps
247131	been	vbn
242854	has	vhz
225582	have	vhb
225381	will	vm0
221172	would	vm0
211159	her	dps
206150	n't	xx0
201616	there	ex0
197483	can	vm0
195515	all	dt0
193757	as	cjs
189926	if	cjs
186984	who	pnq
173582	what	dtq
170417	have	vhi
165805	that	cjt-dt0
161742	that	dt0
160935	said	vvd
159578	its	dps
157972	one	crd
157300	's	vbz
152395	into	prp
151029	him	pnp
150609	some	dt0
148165	could	vm0
140989	them	pnp
138167	when	avq-cjs

134941	time	nn1
129196	out	avp
128980	my	dps
127987	two	crd
127430	up	avp
124543	no	at0
124501	then	av0
123686	more	av0
123315	do	vdb
119113	also	av0
116367	other	aj0
115946	about	prp
112278	these	dt0
110963	me	pnp
108775	first	ord
108669	your	dps
108593	so	av0
108462	did	vdd
108392	new	aj0
108301	now	av0
108088	may	vm0
108043	any	dt0
105560	as	cjs-prp
105411	only	av0
102554	as	prp
102516	people	nn0
101495	than	cjs
100822	her	pnp
99069	should	vm0
87862	like	prp
87705	as	av0
87034	between	prp

86823	very	av0
85826	just	av0

82920	many	dt0
82878	years	nn2

82343	way	nn1

Table 71: The BNC top 100 words of the context-governed spoken corpus

Freq	**word**	**POS**
295636	the	at0
170675	and	cjc
136692	i	pnp
134074	you	pnp
126064	it	pnp
117906	a	at0
117140	of	prf
105313	to	to0
82272	in	prp
75509	's	vbz
75237	we	pnp
70296	is	vbz
67160	er	unc
62789	that	dt0
54810	that	cjt
52865	they	pnp
49147	was	vbd
49078	n't	xx0
48932	to	prp
45807	erm	unc
41895	for	prp
40640	be	vbi
38220	this	dt0
37755	but	cjc
37369	what	dtq
35798	on	prp
31824	are	vbb
31079	do	vdb
29795	if	cjs
29442	with	prp
28845	at	prp
27288	not	xx0
27184	he	pnp

25366	've	vhb
25099	have	vhb
24926	there	ex0
24418	're	vbb
23995	would	vm0
23049	yeah	itj
23033	or	cjc
22778	so	av0
22382	well	av0
21947	yes	itj
21606	can	vm0
21524	that	cjt-dt0
21072	one	crd
20076	just	av0
19464	which	dtq
19348	think	vvb
18841	know	vvb
18541	then	av0
18486	have	vhi
18220	very	av0
17961	all	dt0
17953	were	vbd
17915	now	av0
17734	two	crd
17403	about	prp
17089	from	prp
16711	people	nn0
16679	them	pnp
16116	got	vvn
16107	there	av0
15188	your	dps
14854	because	cjs
14696	did	vdd
14293	do	vdi

14268	an	at0
14247	will	vm0
14171	been	vbn
13500	by	prp
12929	had	vhd
12844	right	av0
12753	some	dt0
12641	up	avp
12596	'll	vm0
12122	could	vm0
12068	going	vvg
12012	'm	vbb
11986	who	pnq
11950	has	vhz
11851	no	itj
11595	my	dps
11569	time	nn1
11541	three	crd
11387	as	cjs
11202	out	avp
11184	how	avq
10891	mm	itj
10821	me	pnp
10809	mean	vvb
10758	oh	itj
10692	get	vvi
10589	their	dps
10431	any	dt0
10195	our	dps
10114	so	cjs
9964	's	pos
9547	when	avq-cjs
9522	actually	av0

Table 72: The BNC top 100 words of the demographic spoken corpus

Freq	**word**	**POS**
167640	i	pnp
135217	you	pnp
128165	it	pnp
115247	the	at0
92239	's	vbz
90886	and	cjc
77611	n't	xx0
68846	a	at0

62382	that	dt0
58810	yeah	itj
48322	he	pnp
47391	to	to0
43977	they	pnp
42241	do	vdb
41654	oh	itj
38515	what	dtq
35156	is	vbz

34901	of	prf
34837	was	vbd
34477	in	prp
33763	she	pnp
33166	we	pnp
31662	no	itj
30177	well	av0
27233	but	cjc
23297	to	prp

22779	've	vhb
22567	for	prp
22016	got	vvn
21907	mm	itj
21586	know	vvb
21400	not	xx0
21370	er	unc
21241	on	prp
20247	did	vdd
19720	're	vbb
19585	this	dt0
19563	just	av0
19212	'll	vm0
18698	be	vbi
18284	there	av0
18051	said	vvd
17898	yes	itj
17809	have	vhb
17610	then	av0
17368	if	cjs
16619	erm	unc
16558	with	prp
16557	are	vbb
16292	have	vhi
15953	so	av0

15746	them	pnp
15367	me	pnp
15297	can	vm0
14477	your	dps
14261	all	dt0
14217	up	avp
14048	'm	vbb
13743	at	prp
13148	that	cjt
12918	there	ex0
12539	get	vvi
12044	my	dps
11952	like	prp
11911	do	vdi
11799	or	cjc
11585	now	av0
11455	one	Crd
11378	cos	cjs
10570	were	vbd
10560	out	avp
10541	think	vvb
10484	go	vvi
10468	mean	vvb
10390	two	crd
10251	going	vvg

10120	know	vvi
10051	na	to0
10021	would	vm0
9573	had	vhd
9163	really	av0
9161	right	av0
8984	one	pni
8896	him	pnp
8812	's	vhz
8769	about	prp
8443	here	av0
8367	how	avq
8100	could	vm0
8087	ca	vm0
8085	gon	vvg
7812	some	dt0
7807	does	vdz
7703	when	avq-cjs
7545	good	aj0
7471	that	cjt-dt0
7461	on	avp
7421	been	vbn
7371	go	vvb
7344	down	avp

The results of Table 73 below are from Leech et al. (2001: 144) since Kilgarriff's website does not have readily available lists for the spoken component of the BNC.

Table 73: The BNC top 100 words of the spoken part of the whole corpus

39605	The	Det
29448	I	Pron
25957	you	Pron
25210	and	Conj
24508	It	Pron
18637	A	Det
17677	's	Verb
14912	to	Inf
14550	of	Prep
14252	that	DetP
12212	n't	Neg
11609	in	Prep
10448	we	Pron
10164	is	Verb
9594	do	Verb
9333	they	Pron
8542	er	Uncl
8097	was	Verb
7890	yeah	Int
7488	have	Verb
7313	what	DetP

7277	he	Pron
7246	that	Conj
6950	to	Prep
6366	but	Conj
6239	for	Prep
6029	erm	Uncl
5790	be	Verb
5659	on	Prep
5627	this	DetP
5550	know	Verb
5310	well	Adv
5067	so	Adv
5052	oh	Int
5025	got	Verb
4735	've	Verb
4693	not	Neg
4663	are	Verb
4544	if	Conj
4446	with	Prep
4388	no	Int
4255	're	Verb

4136	she	Pron
4115	at	Prep
4067	there	Ex
3977	think	Verb
3840	yes	Int
3820	just	Adv
3644	all	DetP
3588	can	VMod
3474	then	Adv
3464	get	Verb
3368	did	Verb
3357	or	Conj
3278	would	VMod
3163	mm	Int
3126	them	Pron
3066	'll	VMod
3034	one	Num
2894	there	Adv
2891	up	Adv
2885	go	Verb
2864	now	Adv

2859	your	Det
2835	had	Verb
2749	were	Verb
2730	about	Prep
2710	two	Num
2685	said	Verb
2532	one	Pron
2512	'm	Verb
2507	see	Verb
2444	me	Pron
2373	very	Adv
2316	out	Adv
2278	my	Det

2255	when	Conj
2250	mean	Verb
2209	right	Adv
2208	which	DetP
2178	from	Prep
2174	going*	Verb
2116	say	Verb
2082	been	Verb
2063	people	NoC
2039	because*	Conj
1986	some	DetP
1949	could	VMod
1890	will	VMod

1888	how	Adv
1849	on	Adv
1846	an	Det
1819	time	NoC
1780	who	Pron
1776	want	Verb
1762	like	Prep
1737	come	Verb
1727	really	Adv
1721	three	Num
1663	by	Prep

Below we compare the top 100 words of the whole corpus (WC*) against the written (WR*) and spoken (SP*) subcorpora of the BNC against the top 100 words of the whole corpus. The results of this comparison are subsequently compared to the results of experiments carried on the Setswana corpus above.

Table 74: Comparison of the top 100 words of the BNC against the top 100 words of the written and spoken subcorpora

N	WC	WR	SP
1	the	1	1
2	of	2	9
3	and	3	4
4	a	4	6
5	in	5	12
6	to	6	-
7	it	10	5
8	is	7	14
9	was	9	18
10	to	8	8
11	i	16	2
12	for	11	26
13	you	20	3
14	he	13	22
15	be	14	28
16	with	12	40
17	on	15	91
18	that	17	23
19	by	18	100
20	at	19	44
21	are	22	38
22	not	23	37
23	this	25	30
24	but	27	25
25	's	21	-
26	they	31	16

27	his	24	-
28	from	26	81
29	had	28	65
30	she	30	43
31	which	29	80
32	or	32	54
33	we	35	13
34	an	33	92
35	n't	43	11
36	's	57	7
37	were	34	66
38	that	53	10
39	been	37	84
40	have	39	20
41	their	36	-
42	has	38	-
43	would	41	55
44	what	50	21
45	will	40	89
46	there	44	45
47	if	48	39
48	can	45	50
49	all	46	49
50	her	42	-
51	as	47	-
52	who	49	94
53	have	51	-

54	do	72	15
55	that	52	23
56	one	56	59
57	said	54	69
58	them	62	57
59	some	60	87
60	could	61	88
61	him	59	-
62	into	58	-
63	its	55	-
64	then	70	51
65	two	67	68
66	when	63	77
67	up	68	61
68	time	64	93
69	my	66	76
70	out	65	75
71	so	80	33
72	did	81	53
73	about	75	67
74	your	74	64
75	now	83	63
76	me	77	73
77	no	69	41
78	more	71	-
79	other	74	-
80	just	97	48

81	these	76	-
82	also	73	-
83	people	89	85
84	any	85	-
85	first	78	-
86	only	87	-
87	new	82	-
88	may	84	-

89	very	96	74
90	should	92	-
91	as	86	-
92	like	93	96
93	her	91	-
94	than	90	-
95	as	88	-
96	how	-	90

97	well	-	32
98	way	100	-
99	our	-	-
100	as	94	-
TOTAL		**97**	**72**

N = Word rank

WC = Whole corpus

WR = Written subcorpus

SP = Spoken subcorpus

Ninety percent of the BNC is written material while 10% is transcribed speech. Ninety seven of the top 100 words of the written component of the corpus are found in the most frequent 100 words of the whole corpus. On the other hand, only 72 words of the top 100 words of the spoken component are found amongst the most frequent 100 words of the entire corpus. Table 75 below shows a comparison of the results of the BNC and of those of the Setswana corpus.

Table 75: Comparison of BNC and Setswana

Corpus	Written Component	Spoken component
BNC	97	72
Setswana Corpus	81	71

When we compare the BNC results with those of the Setswana corpus components we find that 71 of the Setswana spoken subcorpus' most frequent 100 words are found amongst the most frequent 100 Setswana words. Eighty one of most frequent 100 written words are found amongst the most frequent 100 words of the complete Setswana corpus. The results are fairly similar, particularly those of the spoken part of the corpus. The corpus components are also comparable since the BNC has 90% written material and 10% transcribed speech while Setswana corpus is 94% written material and 6% transcribed speech.

Both the top 100 written and spoken components of the corpus do not have all the words found in the top 100 words of the whole corpus. It is however worth noting that

the written and spoken components are complimentary since words which are absent in one subcorpus may be found in another subcorpus.

7.5 A direct comparison of Setswana spoken and written corpus components

Having compared the most frequent 100 words of spoken and written language against the most frequent 100 words of the corpus by seeing which words of each subcorpus are present in the top 100 words of the entire corpus, we now turn to comparing both subcorpus components directly with each other. We use Wordsmith Tools' wordlist program to compare the wordlists directly. This program is exactly the same as the keywords program discussed previously and uses log likelihood statistic as well.

> The procedure compares all the words in both lists and reports on all those which appear significantly more often in one than the other, including those which appear more than a minimum number of times in one even if they do not appear at all in the other (Scott, 2004-2006: 106).

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent to spoken language. At the end of the listing are those words which are outstandingly infrequent in spoken language but are key to written language. In Table 76, we give the most frequent 30 words in spoken language and Table 77 gives the most infrequent words in spoken language, or the most key words in the written language.

Table 76: Outstandingly frequent spoken language

N	Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
1	gore	26,232	3.12	98,272	0.81	28,857.99
2	re	27,313	3.24	130,324	1.08	21,884.16
3	mr	2,625	0.31	515		11,627.94
4	ko	3,528	0.42	3,969	0.03	9,469.37
5	ke	27,070	3.22	214,179	1.77	7,551.83
6	rraetsho	1,592	0.19	257		7,251.76
7	hansard	1,191	0.14	0		6,514.12

8	honourable	1,126	0.13	4		6,105.93
9	jaanong	3,722	0.44	9,556	0.08	5,892.05
10	leng	3,301	0.39	7,355	0.06	5,856.40
11	speaker	1,071	0.13	8		5,764.31
12	the	4,960	0.59	19,603	0.16	5,055.83
13	motsamaisa	991	0.12	85		4,836.79
14	2002	800	0.10	0		4,375.22
15	member	820	0.10	31		4,222.55
16	resumed	756	0.09	0		4,134.54
17	debate	799	0.09	46		4,018.69
18	page	1,331	0.16	1,164		3,988.55
19	dipuisanyo	980	0.12	343		3,891.60
20	ee	2,104	0.25	4,489	0.04	3,854.29
21	palamente	1,124	0.13	896		3,493.26
22	motion	654	0.08	11		3,466.05
23	ba	28,570	3.39	283,076	2.34	3,364.43
24	bill	641	0.08	51		3,148.24
25	of	2,812	0.33	11,011	0.09	2,895.52
26	rona	3,101	0.37	13,185	0.11	2,876.46
27	bua	3,090	0.37	13,985	0.12	2,631.98
28	ra	1,839	0.22	5,575	0.05	2,500.78
29	yone	2,353	0.28	9,187	0.08	2,430.83
30	kana	2,581	0.31	10,927	0.09	2,406.97

Table 76 and 77 are sorted on the basis of keyness or log likelihood statistic listed on the seventh column on the extreme right of the table. RC. Freq. and RC. % refers to the word frequency of the reference corpus and reference corpus's word percentage respectively.

Table 77: Outstandingly infrequent spoken tokens

N	Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
1,954	Yo	2,086	0.25	34,334	0.28	-173,359.36
1,955	gagwe	728	0.09	33,421	0.28	-177,059.83
1,956	nna	3,433	0.41	39,498	0.33	-194,254.70
1,957	la	2,669	0.32	45,661	0.38	-231,733.53
1,958	mme	6,936	0.82	52,649	0.43	-249,051.45
1,959	bo	7,160	0.85	54,427	0.45	-257,615.50
1,960	tla	3,356	0.40	51,181	0.42	-258,048.47
1,961	tse	6,061	0.72	62,008	0.51	-303,327.06
1,962	kwa	3,646	0.43	64,275	0.53	-328,216.78
1,963	i	2,878	0.34	68,621	0.57	-356,843.38
1,964	sa	3,881	0.46	76,545	0.63	-394,662.47
1,965	gore	26,232	3.12	98,272	0.81	-423,485.22
1,966	tsa	4,523	0.54	86,900	0.72	-448,427.16
1,967	wa	3,564	0.42	90,486	0.75	-474,360.16
1,968	ne	4,364	0.52	92,154	0.76	-478,703.91
1,969	di	9,586	1.14	113,319	0.94	-568,445.81
1,970	re	27,313	3.24	130,324	1.08	-590,392.75

1,971	se	9,445	1.12	122,154	1.01	-618,585.81
1,972	fa	11,418	1.36	130,440	1.08	-655,166.81
1,973	ga	10,334	1.23	138,306	1.14	-705,009.69
1,974	mo	11,156	1.33	180,148	1.49	-940,332.00
1,975	ke	27,070	3.22	214,179	1.77	-1,060,745.50
1,976	ya	10,538	1.25	215,238	1.78	-1,149,694.00
1,977	ka	19,592	2.33	268,149	2.21	-1,415,689.88
1,978	ba	28,570	3.39	283,076	2.34	-1,461,408.50
1,979	o	7,328	0.87	320,525	2.65	-1,816,507.13
1,980	le	8,687	1.03	345,885	2.85	-1,968,098.25
1,981	e	30,954	3.68	372,429	3.07	-2,006,141.63
1,982	go	27,937	3.32	385,650	3.18	-2,107,078.75
1,983	a	33,154	3.94	643,503	5.31	-3,936,417.75

A look at Table 76 results shows a high level of parliament terminology as evidenced by the following, *Rraetsho* (Sir) (6), Hansard (7), Honourable (8), Speaker (11), *Motsamaisa Dipuisanyo* (13, 19) (Speaker), Member (15), Resumed (16) and Debate (17). This is expected since spoken text in the corpus is dominated by parliamentary Hansard documents.

"*Gore*" (that) is the most key which scores 28,857.99 on the keyness column and "*A*" is the most outstandingly infrequent with -3,936,417.75 on the keyness column. The top infrequent words include mostly words which are members of the class of closed words such as *Ka* (with) (1,977), *Ya* (of) (1,976), *Go* (1,982) (to), *E* (1,981) (it), *Le* (1,980) (and), *O* (1,979) (he/she), *Ba* (1,978) (they, those), *Gore* (1965) (that), *Gagwe* (1,955) (his/hers). We would expect most of these words to appear high in the spoken subcorpus however an inspection of the most frequent words in the written corpus in Table 68 shows that these words rank high in the written corpus as well. Two matters may be responsible for their showing in the written subcorpus: first, is the size of the written subcorpus which is large compared to the spoken part of the corpus. The written subcorpus is 94% of the whole corpus while the spoken subcorpus is only 6%. Second, the 6% of the spoken subcorpus has a large Hansard section. Hansard material occupies 73% of the whole spoken subcorpus.

The comparison is significant since it reveals that there are distinctions between spoken and written language. This observation is relevant to corpus design since recognition of the distinction of spoken and written and spoken language should influence corpus compilers to sample both written and spoken language for inclusion

in a corpus.

7.6 Comparison of opportunistic and balanced corpora

In the past few experiments we have investigated different text types and through keyword analysis isolated those words which are particular to them. The experiments were intended to test whether different text types contribute distinct words. These findings are relevant to corpus design for lexicography in general, and particularly to this thesis. The recognition that different text types contribute different numbers of words should influence lexicographers compiling dictionaries on the basis of corpus evidence to pay closer attention to corpus design to ensure the broadest coverage possible of text types in a corpus. This is for the reasons that the quality of retrieved information for lexicographic purposes depends on the information input at the stage of corpus construction.

To further test whether text type diversity is crucial to the words selected for inclusion in a dictionary, we compare two 5,000-word list chunks. The first chunk simulates an opportunistic corpus (also called convenience sample (Borin, 2000: 76)) with its text type limitations since it is derived exclusively from prose text. We use the prose text since many readily available text materials in many African languages is of a prose type. The majority of such text would comprise novels.

While in many African languages most readily available text will be prose, in other contexts such text may be newspaper text or web text (see Borin, 2000 and Mair, 1992). For instance 900 million words of Afrikaans texts in the Media24[27] archives could be used as corpus material. Such corpora would be heavily skewed towards a single text type and may not be taken to represent language variability that exists in a speech community. A good illustration of this is MacLeod and Grishman (2000) who report on the creation of two machine readable dictionaries COMLEX Syntax and NOMLEX produced at New York University, in which they used the BNC and the Brown Corpus to which they added a large amount of newspaper text. COMLEX

[27] www.media24.com

contained 7 MB of the Brown Corpus, 27 MB of Wall Street Journal, 30 MB of San Jose Mercury, 29.5 MB of Associate Press text and 1.5 MB miscellaneous selections from the Treebank Literature. They illustrate how an increase in the Brown Corpus (which is generally regarded as balanced) of 1,329% (thus more than thirteen times) resulted in a skewed or inadequate corpus:

> First of all, the make-up of the POS corpus, with its preponderance of newspaper text, skewed the choice of high-frequency verbs. This can be seen by comparing the frequency-ranked list from this corpus with that from Brown, a more balanced corpus. Among the top 50 verbs from our corpus, quite a few (business-related) verbs were not in the top 50 from Brown, including *sell*, *rise*, *buy*, *pay*, and *increase*. In fact, some were not even in the top 750 from Brown, such as *post*, *boost*, *invest*, *value*, and *resign* (MacLeod and Grishman, 2000: 142).

Their results show that media publications such as texts from newspapers and journals mostly available in large quantities if used indiscriminately can skew a corpus (see also Čermák and Křen, 2005). Their experiment therefore offer support to the position that the opportunistic approach to corpus building runs the risk of creating a skewed corpus that does not adequately capture the linguistic rich diversity of a language.

Other researchers have also argued against an opportunistic corpus compilation approach. For instance Biber argues for corpus diversity by pointing that:

> …regardless of the corpus size, a corpus that is systematically selected from a single register cannot be taken to represent the patterns of variation in a language; corpora representing the full range of registers are required. …it is important to design corpora that are representative with respect to both size and diversity. However, given limited resources for a project, representation of diversity is more important for these purposes than representation of size (Biber, 1995: 131).

While we share Biber's position on corpus composition, his argument needs to be tested. To test the text type variability assumption the most frequent 5,000 words were

derived from the prose text and compared with 5,000 words from a variety of text types. The second wordlist of 5,000 words mirrors a balanced corpus, while the first wordlist mirrors an opportunistic one. It was compiled with 500 top keywords from the following 10 text types:

I. Newspaper Text
II. Religious Text
III. Chat-site Text
IV. Hansard Text
V. Poetry Text
VI. Prose text
VII. Politics Text
VIII. Science Text
IX. Call-in Text
X. Business Text

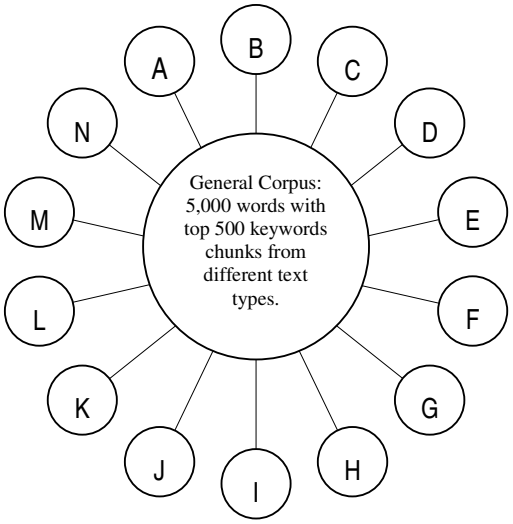
The purpose of comparing the two 5,000-word lists should be by now apparent. It is to measure which of the two lists covers a broad scope of linguistic varieties similar to the one found in the range of varieties of Setswana language use. While we acknowledge that both 5,000-word lists could be used in the compilation of dictionaries, we do however want to measure for wide linguistic coverage in both lists. The question we want to answer is whether the diversity of text types in corpus compilation adds significant value to the quality of dictionary entries by contributing broad word coverage or whether broad word coverage may be attained from a corpus compiled from a single text type, such as prose text.

However, the concept of broad text coverage should not be perceived as restricted to the compilation of corpora for general dictionaries. Even corpora for dictionaries of specialised areas like science and linguistics should demonstrate broad text coverage. This is because specialised areas tend to comprise a variety of even more specialised minute areas. For instance, the area of science is broad; it covers physics, chemistry, biology, engineering, physiology and a variety of other science subjects. The area of linguistics is equally broad comprising morphology, phonology, syntax, semantics, sociolinguistics, psycholinguistics, lexicology, lexicography, computational linguistics and a variety of other areas of language study. Corpora for dictionaries of specialised areas such as linguistics and science should (just like corpora for general dictionaries are compiled with a broad coverage of text types of the general language) also be compiled using a broad range of the text types that constitute the specialised area.

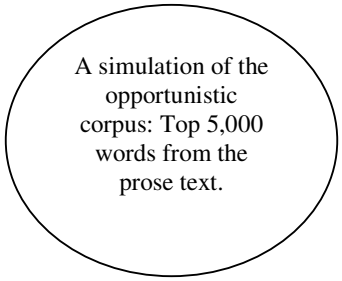
We graphically illustrate how the two 5,000 wordlists are compiled. On one hand is 500-word chunks from different sources compiled together to form a 5,000 wordlist and, on the other hand is the most frequent 5,000 words from a single text type, prose text. From henceforth we will refer to the 5,000 words derived from diverse sources as a combined list.

Figure 17: 5,000 words from a variety of sources

5,000 words from diverse sources



5,000 words from a single source



The 5,000 words from diverse sources were compiled by first sampling the top 500 tokens from each of text types. Each sampled token was sampled with its frequency from its text type. This resulted with overlaps and amongst overlapping tokens, tokens with lower frequencies were deleted and the one with a higher or the highest frequency was kept. We then added 50-tokens incrementally from each of the ten text types, deleting any overlaps until we got 5,000 tokens. Any extra tokens after reaching the 5,000-token were deleted. The tokens were ordered on the basis of their frequencies from their text types. Besides the ordering of tokens within the 5,000 tokens from diverse text types, the frequencies are not used to make comparisons between the two 5,000-token wordlists.

In Table 78 we give the results of the top 100 words from both lists.

Table 78: Top 100 tokens of Prose and Combined list

no	Prose list	Combined list
1.	a	go
2.	go	le
3.	le	ba
4.	e	ka
5.	o	ke
6.	ba	mo
7.	ka	fa
8.	ke	ga
9.	mo	ne
10.	ya	se
11.	fa	i
12.	ga	wa
13.	ne	sa
14.	re	o
15.	se	e
16.	gore	kwa
17.	di	ya
18.	wa	re
19.	sa	gore
20.	kwa	mme
21.	tsa	tla
22.	mme	gagwe
23.	bo	ie
24.	tla	bona
25.	tse	nna
26.	gagwe	the
27.	bona	to
28.	nna	yo
29.	la	fela
30.	yo	posted
31.	fela	ntse
32.	ntse	na
33.	na	itse
34.	itse	tsa
35.	i	neng
36.	neng	and
37.	jaaka	you
38.	motho	motho
39.	lo	di
40.	bone	ene
41.	jwa	on
42.	ene	of
43.	batho	bone
44.	mongwe	at
45.	tswa	by
46.	teng	is
47.	morago	tswa
48.	jalo	that
49.	bua	bo
50.	setse	morago
51.	rile	setse
52.	pele	kwa
53.	monna	tse
54.	eng	rile
55.	thata	pele
56.	ngwana	monna
57.	utlwa	eng
58.	dira	in
59.	nako	ngwana
60.	me	utlwa
61.	batla	are
62.	tsena	nako
63.	gape	it
64.	bana	morena
65.	pelo	batla
66.	letsatsi	tsena
67.	bile	gape
68.	gonne	la
69.	ntlha	pelo
70.	kgotsa	lo
71.	tsamaya	letsatsi
72.	jaanong	bile
73.	tsaya	gonne
74.	kana	gagwe
75.	mosadi	ntlha
76.	gago	tsamaya
77.	sentle	mosadi
78.	wena	sentle
79.	tle	tle
80.	kgosi	for
81.	kae	this
82.	jang	kae
83.	rona	matlho
84.	jaana	not
85.	matlho	botswana
86.	gone	we
87.	sengwe	ena
88.	ena	have
89.	ie	jaanong
90.	dilo	kete
91.	kete	modimo
92.	tota	ise
93.	nngwe	godimo
94.	ise	rata
95.	godimo	gago
96.	rata	twe
97.	bangwe	ko

98.	twe	ja
99.	ja	be

100.	tshwanetse	sepe

At the top of both the prose and combined frequency wordlist are the expected functional words which normally occupy the top rank of frequency lists in a variety of text types. These amongst other words include *a, go, e, le, o, ba, ka, ke, mo, fa, ya, ga, ne, mo* and *se*. The top words are therefore fairly similar to those found in the highest frequency position of the entire corpus. The most frequent words of both lists are therefore not very different from each other save minor differences of various words being at different positions of rank which are not very far from each other. This is positive for both lists since it means that both lists in general capture the most frequent words in the language. For lexicography, it means that if dictionaries were compiled using the two lists, the most frequent words, which in many instances are functional words, would not be excluded from the dictionary.

However to see if the different lists offer significant differences we must inspect the different parts of the two lists preferably looking for the inclusion of words from different text types. We will therefore discuss the inclusion of words in the different lists.

We start looking at religious text. We consider those words which characterise Christianity and traditional Setswana beliefs (TSB). We choose these two since they are followed by the largest percentage of the population with Christianity estimated at 68% and TSB at 30% (Humphries, 2003: 166). We focus on the following words:

Table 79: Christian terms

Setswana terms	English
1. *Jeso*	Jesus
2. *Keresete*	Christ
3. *Modimo*	God
4. *Baebele*	Bible
5. *Bakeresete*	Christians
6. *Legodimo*	Heaven

Table 80: TSB terms

TSB terms	English
1. *Badimo*	ancestors
2. *Moloi*	witch/wizard
3. *Setlhabelo*	sacrifice
4. *Dipheko*	charms
5. *Ditaola*	divination bones
6. *Matwetwe*	traditional doctor

In Tables 81 and Table 82 we offer the results pf the comparisons by showing the rank the words occupy in Prose and combined lists. A discussion of the results follows their presentation.

Table 81: Christian terms and their ranks on the two lists

Setswana	English	Prose text	Combined text
Jeso	Jesus	-	1,222
Baebele	Bible	-	1,303
Bakeresete	Christians	-	2,698
Legodimo	heaven	1,340	1,065
Keresete	Christ	4,855	678
Modimo	God	219	91

Table 82: TSB terms and their ranks on the two lists

Setswana	English	Prose text	Combined text
Badimo	Ancestors/gods	360	-
Moloi	Witch/wizard	1,701	3,319
Setlhabelo	Sacrifice	-	834
Dipheko	Charms	1,944	1731
Ditaola	Divination bones	3,277	2,756
Matwetwe	Traditional doctor	-	3,012

The constant result in both tables is that the combined text numbers are ranked higher in the list compared to the prose words save for *moloi* (witch/wizard) which appears higher in prose text. Some of the gaps between words in the two lists are significantly higher. For instance, the difference in *Keresete* (Christ) is at 4177, *Moloi* (witch/wizard) 1618 and *ditaola* (divination bones) at 521. Second, *Jeso* (Jesus), *Baebele* (Bible) and *Bakeresete* (Christians) do not make it into the top 5,000 prose text. A look at the TSB terms also reveals that *badimo* does not make it into the top

5,000 words of the combined text while *setlhabelo* and *matwetwe* do not make it to the prose text. Therefore in the 12 words that we have inspected in the two tables, 5 of the words do not make it to the top 5,000 prose text and only one does not make it to the top 5,000 combined list. These results are significant in that they reveal that almost half of the inspected words do not make it into the top 5,000 words of prose text.

Lexicographically, the implications are serious. Missing words in a dictionary such as the ones inspected above leads to gaps in the lexical representation of a language in a dictionary.

We now look at the grammar text and inspect some basic grammatical terms and measure the performance of both lists. Grammar texts are studied since they are central to students' Setswana grammar classes which are compulsory at both junior and senior secondary schools. Basic grammatical terms would therefore be expected in school dictionaries, even short ones. Below we present the results of six grammatical terms.

Table 83: Grammar terms and their position on the two lists

Setswana	English	Prose text	Combined text
Tumanosi/ditumanosi	Vowel(s)	-	35911389
Lediri	Verb	884	898
Tumammogo/ditumammogo	Consonant(s)	-	43502247
Letlhaodi	Adjective	4,544	3089
Letlhalosi	Adverb	3,409	2125
Leemedi	Pronoun	-	4569

The results above show that half of the words do not show up in the most frequent 5,000 words of prose text. These are *tumanosi* (vowel) and its plural *ditumanosi* (vowels), *tumammogo* (consonant) and its plural *ditumammogo* (consonants) and *leemedi* (pronoun). The results reveal a lack of some of the basic grammatical labels in the prose text. On the other hand, all the words inspected appear in the combined list. These results are consistent with the previous results where about half of the words do not appear in the restricted list but do appear in the combined list. The results show that while prose texts deal with a variety of subjects they have limitations

when specialised areas like linguistics are studied.

We also look at the business terms and how they perform in both the prose and combined list. The results follow in Table 84:

Table 84: Business terms and their rank on the two lists

Setswana	English	Prose text	Combined text
Bagwebi	Business people	-	787
Kgwebo/dikgwebo	Business/businesses	1151/3978	585/1094
Mmaraka	(market	-	1133
Madirelo	Factories	3935	933
Kompone	Company	-	629
Itsholelo	Economy	-	679

Only two of the six business terms make it into the top 5,000 prose words. The two words that do make it into the top 5,000 are comparatively ranked lower in the list. The business terms results are consistent with the results which have been seen so far with grammar and religious terms from Christianity and traditional Setswana beliefs.

For our final measurement we look at taboo words; insults or vulgarities which rarely make it into school textbooks, local newspapers and dictionaries. Landau (1989: 187) laments that, "[n]o aspect of usage has been more neglected by linguists and lexicographers than that of insults." Their lack of inclusion in such texts is barely surprising since insults are not just taboo, but by their nature they constitute what Butler (1997: 2) calls "injurious speech" or signs used with the intention to shock, bring offence and psychological harm to the targeted individual or group. Insults are therefore instances of linguistic violence; reflections of how individuals verbally inflict injury on each other (cf. McEnery and Xiao, 2003). They take different forms. Some refer to private parts while others are rude words which refer to embarrassing actions particularly when mentioned in public (also see Lynch, 2004: 640[28]). These actions may include references to relieving oneself or they may take the form of coarse words referring to sexual activity or farting. They may also be group insults

[28] Lynch lists insults from Johnson's dictionary which include amongst others: *airling, asshead, backbiter, backfriend, barbarian, bedpresser, bellygod, bitch, blockhead, blowze, blunderhead, booby, barachio, bufflehead, bumpkin, bungler, simpleton, noodle* and *smellfeast*

which refer to and label particular ethnic groups or a particular sex.

There are a variety of reasons why such language is relevant to academic study and more so to have a place in dictionaries. Such reasons include amongst others:

i. Adult learners of a new language, or those who have moved into a new society with a totally different language, may take a keen interest in knowing rude words as a defence mechanism – so that they may be aware when insults are hurled at them.
ii. Insults are taboo; therefore an understanding of insults will contribute to an understanding of a society's taboos – an understanding of what is socially acceptable or profane.
iii. As stated previously, profanities may be perceived as cases of linguistic violence – inflictions of injury on the other. In this way a study of insults may be seen as a study of social violence.
iv. Users search for insults in dictionaries. De Schryver and Joffe in their study that makes a determination of how electronic dictionaries are used. They have found out that,

> [i]n the top 100 searches there are a further 6 foreign words lows (4 Setswana and 2 English), and of the remaining 31 words no less than 17 either have to do with the sexual sphere or are extremely offensive: *marêtê* 'testicles', *masepa* '(*off.*) shit', *mogwêtê* '(*off.*) anus', *mpopo* '(*off.*) private part (vagina; penis)', *nnyô* 'vagina', *nnywana* '(*off.*) cunt', *ntoto* 'penis', *nyôba* '(vulgar) fuck', *sefêbê* 'prostitute; (off.) bitch', *thôbalanô* 'sex', etc. This latter phenomenon might very well be the case for all (Internet) dictionaries (De Schryver and Joffe, 2004: 190).

They also observe that,

> An analogous study of the top 100 English searches reveals a similar pattern, with 18 of the top 100 searches also in the BNC top 100 (Leech et al. 2001) and 62 in the BNC top 1 000. A single item in the

top 100 searches is misspelled, while 6 of the remaining 37 searches again belong to the same sexual/offensive sphere: *bitch*, *fuck, penis, sex, shit* and *vagina* (De Schryver and Joffe, 2004: 190).

De Schryver and Joffe's findings give support to the study of insults as an interesting academic area of investigation.

However the point of this section is not an attempt at a study of profanities but rather to use the absence or presence of insults as an illustration of the strength or weakness of a corpus text type coverage. The point is that if corpora are based on texts which have been edited by publishers and newspaper editors who may be following prescriptive rules about a language, then a corpus itself may be only offering a partial reflection of the state of a language. That is why in the design and compilation of the Setswana language corpus we have incorporated chat-site material. While the material has greater levels of English words, it does provide valuable Setswana language style that is rarely seen in published texts but is characteristic of youthful dialogues. Vulgarities and words that refer to private parts while frequently avoided by publishers do occur in chat-site material.

We therefore look at the different vulgarities and present the results in Table 85:

Table 85: Vulgarities and their position on the two lists

Setswana	English	Prose text	Combined text
Marete	balls, scrotum	-	2269
Polo	dick, penis	-	2087
Masepa	shit	-	1725
Nnyo	vagina	-	982
Phona	vagina	-	-
Sebono	asshole, anus	-	2115

Prose text does not provide any evidence of any of the common vulgarities that we have isolated. This is barely surprising since most of the Setswana prose is primary school, secondary school and university educational material which sanctions vulgarities. Combined texts on the other hand show very high presence of vulgarities.

The different experiments above in which we compare the 5,000 word prose list and

5,000 word combined list aimed at comparing the performance of an opportunistic corpus against a broad coverage corpus. The results of all the experiments point to the inadequacy of an opportunistic corpus (in this case, a single text type corpus) as a reliable source of dictionary material. They reveal that the simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus.

7.7 Chapter conclusion

In this chapter we have explored a variety of experiments to determine if a corpus comprising a variety of text types was any different from one with a single text type on the basis of the types it contributed at comparable intervals. We began by first segmenting the Setswana corpus into 10,000 token chunks. For every text type, the types' measurements were taken at 10,000 token intervals. The 10,000 token-chunks were randomised for every measurement taken and the experiment iterated five times. The type measurements were taken at every 10,000 token interval up to 500,000 tokens. An average was computed so that comparisons between text types using a single mark that summarises the results at every 10,000 tokens interval could be made.

The experiments revealed Poetry text as having the largest number of types at most of the 10,000 tokens intervals followed by PRONEWHANCAL and SCIPOLBUSREL. We have also found out that the combination of text from a variety of text types compiled into POEGRACHAPLA, SCIPOLBUSREL and PRONEWHANCAL resulted with higher types when compared to the distinct text types from which their parts were compiled.

The experiments also revealed that Politics text had the lowest types overall, followed by Call-in and Chat-site texts. This suggests that these three use a limited vocabulary when compared with other text types. We argued that while certain text types contribute the lowest number of types, such a smaller number of types should not be perceived as implying less importance or less significance in corpus compilation, since even the text types with the lowest number of types do contribute unique words to other text types.

The performance of the most frequent 100 words from different text types was measured against the most frequent 100 words of the whole Setswana corpus. It was found out that it was not enough to just have a corpus with a variety of text types to generate large numbers of types. It was also crucial that the individual text types that comprise a corpus should individually have large numbers of word types.

Simple consistency analysis (SCA) which calculates dispersion or word-spread in corpora was also explored. The SCA results were compared to the calculation of raw frequencies in the calculation of the most frequent words in the corpus. SCA has been able to determine whether a widely spread use of a word is because it occurs in many text samples or whether it is frequent because of high usage in only a few texts. The SCA calculation computes words which recur consistently in texts and orders them on the basis of their spread across documents.

The most frequent words in different text types were compared. Raw frequencies were chosen in the comparison of the most frequent 100 words so that the results could be comparable to those of other wordlists of other corpora such as the BNC.

We also compared the most frequent 100 words of the written component of the BNC and the most frequent 100 words of the spoken BNC component against the most frequent 100 words of the whole BNC. The results of this experiment were compared with the Setswana corpus experiment. Seventy one of the Setswana spoken subcorpus' most frequent 100 words were found amongst the most frequent 100 Setswana words. Eighty one of the most frequent 100 written words were found amongst the most frequent 100 words of the complete Setswana corpus. The BNC, on the other hand had 97 of the top 100 words of the written component of the corpus in the most frequent 100 words of the whole corpus and 72 words of the top 100 words of the spoken component. The Setswana corpus therefore compared well with the BNC corpus in this experiment.

To further test whether text type diversity was crucial to the kind of words which are selected for inclusion in a dictionary, two 5,000-word list chunks were compared. The first chunk simulated an opportunistic corpus with its text type limitations since it was

derived exclusively from the Prose text. Prose text was chosen since many readily available text materials in African languages are of a prose type which would comprise novels. The most frequent 5,000 words were therefore derived from the prose text and compared with 5,000 words compiled from a variety of text types. Both lists were tested for a variety of grammatical, religious and business terms and for certain Setswana vulgarities. The results showed that the simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus. It was also found out that some of the most frequent words in Setswana were found in both corpora. The results provide evidence for broad text type coverage in corpora compilation as a reliable source of broad lexical coverage for dictionary compilation.

What the different experiments have shown is that there are considerable differences between the different wordlists extracted from the diverse text types. The experiments testify to the limitation of a single text types as a source of dictionary evidence. They have shown that to get a variety of words of a language, a corpus with diverse text types is preferable.

Chapter 8

Conclusion and future work

This research set out to try and determine how a Setswana corpus could be compiled and structured as a balanced and representative entity through both quantitative and qualitative means in order for it to be "better suited" for lexicography. The aim was to determine whether a corpus compiled with texts from various text types was better suited for lexicography or whether similar results could be attained through corpora compiled with texts from few or a single text type.

To test the aims, a Setswana corpus of over 13 million tokens was compiled. The compilation of this corpus has been discussed in Chapter 5.

Chapter 1 positioned this thesis squarely within the scope of corpus design with specific application to the lexicography of the Setswana language. The chapter outlined the goals and aims of the study, and the methodologies employed. The chapter concludes by taking a panoramic overview of the whole study with an exposition of chapters.

Chapter 2 explored the Setswana language situation particularly in Botswana. Multilingualism in Botswana was discussed and it was shown that Setswana is spoken by the majority of the Botswana population (about 80%). Despite the population's multicultural composition, it has been established that only two languages, Setswana and English, occupy a dominant position in the educational system (Mooko, 2004: 181/2). It was also established that English remains the official language and a language of considerable prestige, while Setswana is the national language and the country's lingua franca. Other Botswana languages apart from Setswana and English have no official status in Botswana (Molosiwa, 2004: 6) and remain excluded from functioning as mediums of instruction and use in the media (both broadcast and print, save for Ikalanga which is used minimally in *Mmegi* Newspaper insert, *Naledi*), parliament, or in any public domain to communicate government policy. It has been shown that while minority languages are in general marginalised from any official

function, in regions where minority languages are the regionally dominant languages, the minority language is usually used in official roles, like communicating with the chief or nurse (Hasselbring et. al. 2001: 32-33). The appraisal of the language situation in Botswana mapped out the complex language use and varieties of Setswana.

We also traced the history of Setswana lexicography and language research dating to the early missionary period and situated them within missionary literacy programs amongst the Batswana. Developments in corpus and computational models have been reviewed and it has been shown how they have affected dictionary compilation in Setswana. We have illustrated how the Setswana language could benefit from developments in corpora, corpus querying software (CQS) to produce frequency lists, concordances, and keyword analysis. By outlining Botswana's sociolinguistic situation, Chapter 2 established a foundation for Chapter 5, which outlines the design and compilation of the Setswana corpus.

Chapter 3 explored the theoretical issues related to corpus design. It surveyed various definitions of "corpus". The following findings were established in the definitions discussed:

- Corpora are usually "sufficiently large" for the research they have been compiled for.
- Corpora are collections of running texts. They are not just lists of words but rather chunks of texts like chapters of books, entire books, or transcribed speech.
- Corpora are compiled for some linguistic research.
- Because of their massive size, corpora are stored in computers because of the computer's storage and processing power. Computers are "good at recall, people are good at precision; that is, computers are good at finding a large set of possibilities, people are good judges of which possibilities are appropriate" (Kilgarriff, 2003: 1). They can also be used interactively, allowing the human analyst to make difficult linguistic judgements.

It was also shown how the Web, with its billions of words, has revolutionised the compilation of corpora. The Web was seen as providing a cheap route to corpus compilation (De Schryver, 2002). The benefits of Web text were revisited and demonstrated in Chapter 5 which details the compilation of half a million tokens through the use of a Web crawler.

Keyness and frequency profiling were introduced in Chapter 3 and later used in Chapter 6 and 7. Baroni (2006: 1) has observed that "The frequency of words and other linguistic units play a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language."

In Chapter 4 we discussed issues which arise in corpus design as they relate to lexicography. Corpus design is at the heart of this study since corpus design and compilation determine the quality of what could be extracted from a corpus. Balance and representativeness have been discussed and found to still be areas of great contestation in corpus design and compilation. It has also been shown that what constitutes balanced and representative corpora still remains controversial. The sampling of genre quantities for a corpus is still largely unresolved. The general consensus in the literature is that a corpus must capture the language varieties of a population from which a sample is taken, which reflects how that particular language community uses language. Such a goal is important since many corpus linguists hope to generalise the results of corpora analysis to the general language community from which the samples have been abstracted.

Sinclair (2004) has argued that the complicating factor in compiling balanced and representative corpora is that language is an "unlimitable phenomena". A quest to quantify language usually results in general estimates. Although language is an unlimitable phenomenon, that has not obstructed corpus researchers from arguing for sampling different linguistic varieties for both quantitative and qualitative study. The challenge for corpus linguists and lexicographers is to identify the language varieties of the language under study and ensure that they are represented in a corpus.

It has also been demonstrated that a corpus can contain simple raw text or it can be enriched with linguistic information which will enhance information extraction. Tagged corpora, it was argued, are useful in the development of disambiguation rules and the facilitation of automatic and semi-automatic syntactic analysis in corpus linguistic. Tagged corpora have also been found to be highly useful in the generation of Word Sketches "…one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004: 105).

In Chapter 4 the importance of spoken language in corpus design and compilation is discussed. Sinclair (2004) shows that "estimates of the optimal proportion of spoken

language range from 50% — the neutral option — to 90%, following a guess that most people experience many times as much speech as writing." However in none of the large corpora like the LOB, Brown Corpus, BNC and the Bank of English does the percentage of spoken text exceed that of written text. This state of corpora has been criticised by Sinclair (2004) thus: "most general corpora of today are badly balanced because they do not have nearly enough spoken language in them."

In the debate on spoken language inclusion in a corpus, Biber (1994) has argued that to have greater percentages of spoken language in a corpus is not linguistically interesting since it leads to corpus homogeneity. He has argued that what corpus compilers should aim for should be stratified corpora that capture the linguistic variability of the language community and not proportionally-compiled corpora.

Chapter 4 also discussed the importance of spoken language by illustrating what would be lost if a corpus lacked transcribed speech. Such losses would include borrowings and code-switching which are linguistically interesting. Setswana data has revealed that spoken Setswana has high concentrations of borrowings from English and Afrikaans and instances of code-switching. Chapter 4 concluded by reviewing the BNC and the Brown corpus. Their internal structures were studied to reveal how they compare with the Setswana corpus compiled for this study. Both corpora include samples from different domains to attempt a representativeness of the English language as used. Their analysis provided a base for Chapter 5 where we outlined the design and compilation of the Setswana corpus.

Chapter 5 mapped out the compilation of the Setswana corpus which has been used for experiments in this study. It is about 13 million words and covers texts from different varieties of Setswana language use including, novels, plays, newspapers, grammar books, spoken language covering court transcripts, call-in programs, television debates, funeral services, classroom interaction and sermons. The corpus, like many corpora, has large sections of written language and smaller sections of transcribed spoken language. Ninety four percent of the corpus is the written component while the spoken component occupies 6%. Its design and compilation was influenced largely by the structure of the BNC. The corpus is significant for the experiments which have been conducted in this thesis as Dash and Chaudhuri (2000: 188) have observed that "the potentiality of a well-designed corpus is immense as it provides an empirical basis for language description"

Keywords for Science and Technology, Politics, Poetry, Plays, Grammar, Arts and

Culture, Religious and Hansard text and interviews text from spoken language were calculated in Chapter 6. The statistical analyses were conducted through the use of WordSmith Tools. The Log Likelihood (G2) test was used to calculate keyness since it is considered better than the chi-square test of significance particularly when contrasting long texts or where one may have to deal with low counts of less than 5. We were here following Kilgarriff (2001: 105) who argued that "G2 is a mathematically well-grounded and accurate measure of surprisingness," and that "it corresponds reasonably well to human judgments of distinctiveness." The results of the experiment were presented as the top 100 keywords from each text type. Through keyword analysis it was found that different text types generated different keywords that were particular to them. Thus terms key to the following texts were abstracted: Religious, Science and technology, Politics, Poetry, Plays, Grammar, Arts and culture, Chat-site, News, Sport, Call-in, Face to face dialogue, Education, Hansard, and Interview and Open-radio programming.

The results of the experiments revealed that different text types contribute different keywords that are unique to them. The finding is significant to corpus design for it lends support to the inclusion of texts from a variety of text types in a corpus. Since text types are characterised by unique words which are key to them, a corpus comprising texts from the varieties of a language will be richer in its representation of a language.

Chapter 7 measured how for each text type the number of types grew with every additional 10,000 tokens. Our aim with this experiment was to investigate whether different text types' vocabularies vary at comparable token points. It was found out that taking measurements at 10,000 token-chunk intervals is sensitive to the order in which texts (i.e. 10,000-word corpus chunks) are placed or ordered. The ordering of the 10,000 token-chunks raised unique challenges in that every experiment repetition with a random order of the 10,000 token chunks gave different results dependent on which 10,000 token-chunk was analysed first.

The 10,000 token-chunks were randomised for every measurement taken and the experiment iterated five times to resolve the bias of sequence. A mean was computed so that comparisons could be made between text types using an average that summarised the results.

For additional experiments the written part of the Setswana corpus was divided into 10 text types of Poetry, Grammar, Chat-site, Plays, Prose, Science, Politics, Business,

Religious, and Newspaper texts. The spoken subcorpus was segmented into two parts: Hansard and Call-in (comprising interview, call-in text and open-radio programming treated as a single unit). Additionally, the 12 text types of spoken and written text were divided into three random groups of A, B and C with each group having four text types. These were labelled using the initial three letters of each text type found in each group, thus: POEGRACHAPLA (Poetry, Grammar, Chat-site and Plays), PRONEWHANCAL (Prose, Newspaper Hansard and Call-in) and SCIPOLBUSREL (Science, Politics, Business and Religious). The experiment compared subcorpora containing unrelated texts with equal-sized subcorpora containing text from a single genre. The TTR measure at comparable points for both texts was computed. The assumption was that combining text from a variety of sources (as one might do in corpus compilation) would give a higher TTR at comparable points compared to that of an equal-sized subcorpus with a single text type.

Randomly sampled fifty 10,000 token-chunks were taken from the 4 text types in each group. The final number of text types came to 15 including the three "constructed text types." Measuring word types of text types at 10,000 tokens intervals revealed that Poetry, PRONEWHANCAL and SCIPOLBUSREL had the largest number of word types in all text types measured. It was concluded that the high levels of types in poetry offer support to the view that poetic language is characterised by high lexical density. The high levels of types in PRONEWHANCAL and SCIPOLBUSREL led to the conclusion that corpora comprising a variety of text types have a higher number of types, than those compiled from a single text type.

The most frequent 100 words of various text types were compared with the top 100 words of the whole corpus. It was found that it was not enough to have a corpus which has a large variety of text types for one to generate large numbers of word types. Rather, it was crucial that the individual text types that comprise a corpus should individually have higher levels of word types themselves.

Simple consistency analysis (SCA) which calculates dispersion or word-spread in corpora was explored in frequency analysis. SCA demonstrated whether a high frequency word was high in the frequency list because it occurred in many of the text samples or whether it was because it was used frequently in a few texts. The SCA calculation computes words which recur consistently in texts and orders them on the basis of their spread. SCA was compared to raw frequency calculations of the most frequent words in the corpus. It was found that the compilation of headword lists would be enhanced by combining SCA calculations and raw frequency lists.

Raw frequencies were then used in the comparison of the most frequent 100 words of different text types so that our results could be comparable to those of other wordlists of other corpora such as the BNC. The most frequent 100 words of the spoken text and the most frequent 100 words of the written part of the corpus were compared with the most frequent 100 words of the entire corpus. It was found that 81 of the top 100 words of the written component of the corpus are found amongst the most frequent 100 words of the whole corpus. On the other hand, only 71 words of the top 100 words of the spoken component were found amongst the most frequent 100 words in the entire corpus. The written component of the corpus is larger and is much more diverse in the kind of texts it comprises while the spoken component is smaller and limited in its text diversity. This may explain the differences between the results of the spoken and written corpus component. Although 94% of the whole corpus is written material, still all of the top 100 in the whole corpus are not found amongst the most frequent 100 words of the written component. This fact suggests that even with its great diversity and size, written language alone is not adequate to make a representative corpus. There is a need for spoken material to be included in the corpus.

The Setswana experiment results of the written and spoken language were compared to those of experiments on the BNC. It was found that 71 of the Setswana spoken subcorpus' most frequent 100 words were amongst the most frequent 100 Setswana words. Eighty one of most frequent 100 written words were found amongst the most frequent 100 words of the complete Setswana corpus. The BNC, on the other hand had 97 of the top 100 words of the written component of the corpus in the most frequent 100 words of the whole corpus and 72 words of the top 100 words of the spoken component. The results of the two experiments were found to be similar and the Setswana corpus components comparable to those of the BNC since the BNC has 90% written material and 10% transcribed speech while the Setswana corpus has 94% written material and 6% transcribed speech.

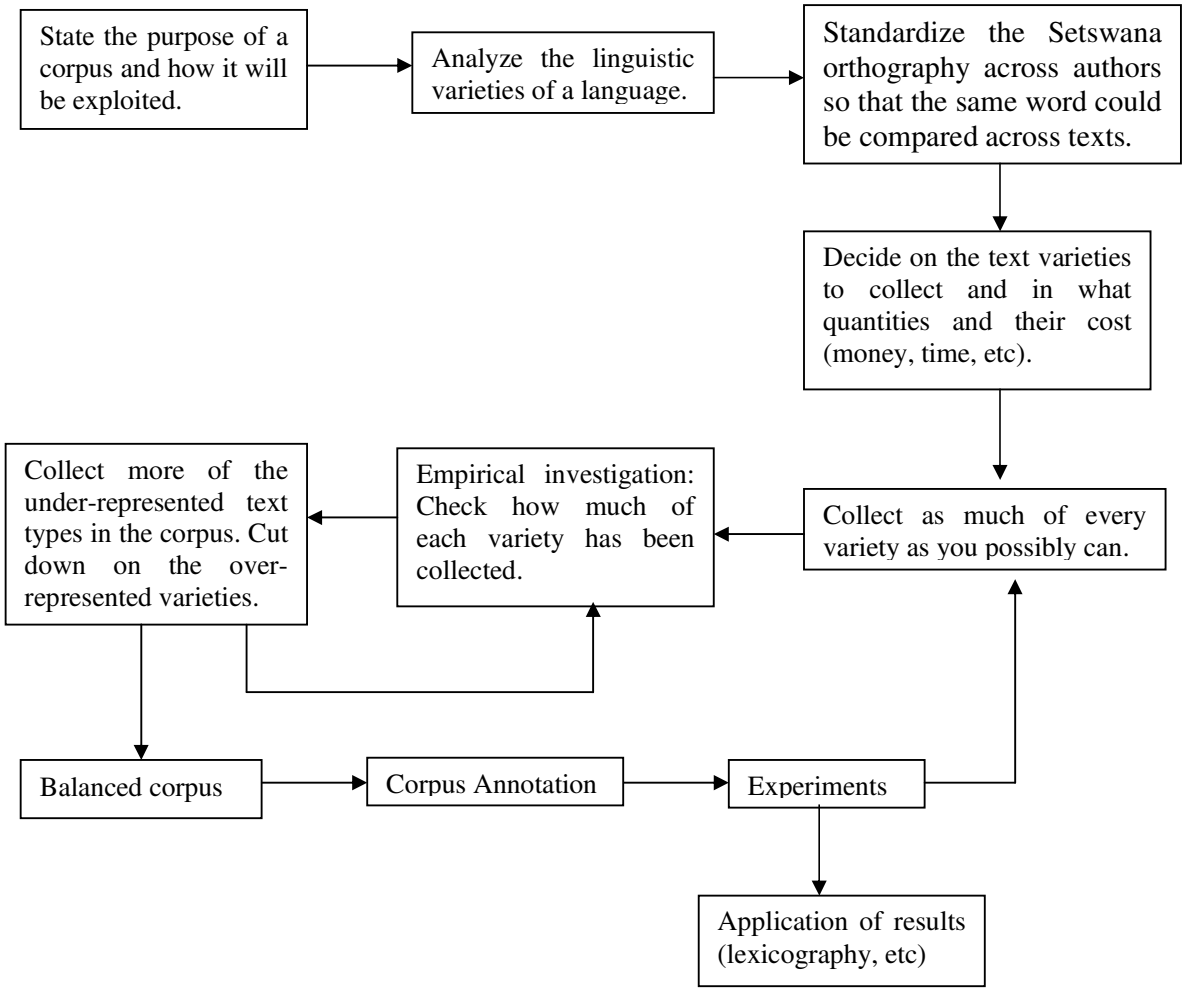
To further test whether text type diversity was crucial to the kind of words selected for inclusion in a dictionary, two 5,000-word list chunks were compared. The first chunk simulated an opportunistic corpus with its text type limitations. It was derived exclusively from prose text. Prose text was chosen since much of the readily available text in African languages is of a prose type. The majority of such text comprises novels. The most frequent 5,000 words from prose text were compared with 5,000 words from the following text types: Newspaper text, Religious, Chat-site, Hansard,

Poetry, Prose, Politics, Science, Call-in and Business. The comparison of the two 5,000-word lists was to determine which of the two lists covered a broad scope of linguistic varieties similar to those found in Setswana varieties. Christian terms, terms common in traditional Setswana beliefs, grammatical terminology, business terms and vulgarities were tested for in the two wordlists.

The results of all the experiments consistently showed that the 5,000 wordlist compiled from a diversity of text types performed better that the 5,000 wordlist from prose text. The 5,000 wordlist compiled from a diversity of text types was found to have a large number of Christian terms, words common in traditional Setswana beliefs, grammatical terms, business words and vulgarities. The results suggest that an opportunistic corpus is largely unreliable as a source of dictionary material. The simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus. The results therefore give support to broad text type coverage in corpora compilation as a reliable source of broad lexical coverage in dictionary compilation.

The experiments conducted in both Chapter 6 and 7 provide an overwhelming support for a diversity of text types in corpus compilation. The implications for corpus design are that stages that can be followed for a model development of a Setswana corpus can be proposed. The same model may be extended to languages similar to Setswana. Below we develop a corpus compilation schema which has been influenced by Biber's schematic representation of the corpus construction (Biber 1994: 400) and by the experiences drawn from this study. The schematic representation illustrates how the purpose to which a corpus is created and an understanding of language varieties that exist in a language inform the corpus model that guides the construction of balanced corpora. The schema is rendered in Figure 18.

Figure 18: Schematic representation of a balanced corpus construction



The schema portrays corpus compilation as a continuous process of collection of text and attempting to balance the different text types against each other. This schema shows that as more text is collected more balancing should be attempted. The collection of more text and subsequent corpus balancing should however not curtail the use of the corpus since more text should ideally be continuously collected particularly for languages like Setswana where texts are generally rare.

8.1 Future research and applications

This study has attempted to contribute to the body of research in corpus design for the Setswana language and languages in a similar position to Setswana. More has to be investigated concerning Setswana corpus design and use on language similar to Setswana. It was not this thesis' aim to attempt to cover everything in corpus design,

such a goal is unattainable. It is hoped that this study will generate debate and research on design and corpus use. More work still has to go into the POS tagging of Setswana corpus to maximise its exploitation. The corpus exploitation discussed in Chapter 4, such as in the case of Word Sketches reported by Kilgarriff (2004) would benefit Setswana research if the corpus was tagged.

Although we have built a 13 million corpus with a variety of text types for this thesis, more text still needs to be collected particularly, spoken language with its diverse varieties. The recording of Setswana dialects for inclusion in the corpus may also shed light on the different terms and sentence structures particular to regional dialects.

It is hoped that this study opens new doors to increased corpora study of Setswana. Genre studies on the basis of corpora evidence have not been attempted in the Setswana language, largely because of the lack of Setswana corpora that comprise different genres. Keyword analysis of text types in Chapter 6 and other experiments in Chapter 7 that explored the text type differences have resulted in fruitful findings which lexicographers, sociolinguistics and others linguists could benefit from. It is also hoped that the methodologies employed and findings of this research will all prove fruitful to other language researchers.

Dash and Chaudhuri (2000: 180) observe that "[p]eople in every branch of information science now realize that a corpus, as a sample of living language, can open up new horizons of study and research." It is therefore hoped that the 13 million Setswana corpus compiled during this study will be a resource for corpora investigation of different aspects of Setswana research beyond this study such as morphology, syntax and further investigations of text type variability.

Bibliography

Aarts, J, and Meijs, W, (eds.). 1990. *Theory and Practice in Corpus Linguistics*. Vol. 4, *Language and Computers: studies in practical linguistics*. Amsterdam: Rodopi.

Aitchison, J. 1992. Teach yourself linguistics. London. Hodder & Stoughton.

Algeo, J. 1988. British and American grammatical differences. *International Journal of Lexicography* 1 (1):1-31.

Alidou, H. 2004. Medium of instruction in post-colonial Africa. In *Medium of Instruction Policies: Which agenda? Whose Agenda?*, edited by J. W. Tollefson and A. B. M. Tsui. Mahwah: Lawrence Erlbaum Associates.

Allwood, J. 1998. Some frequency based differences between spoken and written Swedish. Paper read at The xvi:th Scandinavian conference of Linguistics, at University of Turku.

Al-Sulaiti, L. 2004. Designing and Developing a Corpus of Contemporary Arabic. MSc. thesis, School of Computing, University of Leeds, Leeds.

Andersson, L-G, and Janson, T. 1997. *Languages in Botswana; language ecology in Southern Africa*. Gaborone: Longman Botswana.

Archbell, J. 1837. *A grammar of the Bechuana language*. Cape of Good Hope: Meurant and Godlonton.

Arthur, J. 1997. 'There must be something undiscovered which prevents us from doing our work well': Botswana primary teachers' views on educational language policy. *Language and education* 11 (4):225-241.

Arua, A.E, and Magocha, K. 2002. Patterns of language use and language preference of some children and their parents in Botswana. *Journal of multilingual and multilingual development* 23 (6):449-461.

Aston, G. 2001. Text categories and corpus users: A response to David Lee. *Language Learning and Technology* 5 (3):73-76.

Aston, G, and Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. T. McEnery and A. Wilson (Eds), *Edinburgh Textbooks in Empirical Linguistics*. Edinburgh: Edinburgh University Press.

Atkins, S., Clear, J. and Ostler, N. 1992. Corpus Design Criteria. *Journal of Literary and Linguistic Computing* 7 (1):1-16.

Atkins, S., Kilgarriff, A., and Rundell, M. 2001. Training workshop in lexicography and lexical computing: course notes. Brighton: University of Brighton.

Baayen, R.H. 2001. *Word Frequencies Distribution*. London: Kluwer Academic Publishers.

Bagwasi, M.M. 2003. The Functional Distribution of Setswana and English in Botswana. *Language, Culture and Curriculum* 16 (2):212-217.

Barnbrook, G. 1996. *Language and Computers; A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Baroni, M. 2006. *Distributions in text.* In Lüdeling, A and Kytö, M. (eds.), Corpus linguistics: An international handbook, Berlin: Mouton de Gruyter.

Baroni, B. and Ueyama, M. 2006. Building general- and special-purpose corpora by Web crawling. Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application. 31-40.

Baugh, S., Harley, A., and Jellis, S. 1996. The role of corpora in compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics* 1 (1):39-59.

Béjoint, H. 2000. *Modern lexicography: An introduction*. New York: Oxford University Press.

Bentivogli, L., Girardi, C. and Piata, E. 2003. The MEANING Italian Corpus. Paper read at the Corpus Linguistics 2003 Conference, Lancaster, UK.

Bentivogli, L. and Pianta, E. 2002. Detecting hidden multiwords in bilingual dictionaries. Paper read at The 10th EURALEX International Conference, at Copenhagen, Denmark.

Bergenholtz, H, and Tarp, S. (eds). 1995. *Manual of specialized lexicography*. Amsterdam, Philadelphia: John Benjamins.

Bharati, A., Prakash, R.K., Sangal R, Bendre, S.M. 2002. Basic statistical analysis of corpus and cross comparison among corpora. Proceedings of International Conference on Natural Language Processing, Mumbai, India. 18-21, December 2002. Available at: http://www.iiit.net/techreports/2002_4.pdf.

Biber, D. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5 (4):257-269.

Biber, D. 1993. Using register-diversified corpora for general language studies. *Association for Computational Linguistics* 19 (2):219-241.

Biber, D. 1993a. The multi-dimesional approach to linguistic analyses of genre variation; an overview of methodology and findings. *Computers and Humanities* 26:331-345.

Biber, D. 1994. An analytical for framework register studies. In D. Biber and E. Finegan (eds) *Sociolinguistic Perspetives on Register*, Oxford: Oxford University Press.

Biber, D. 1994a. Representative in Corpus Design. *Linguistica Computazionale: Current Issues in Computational Linguistics: In Honour of Don Walker* ix (x):377-407.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison,*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen R. 1998. *Corpus Linguistics: Investigating Language Structure and Usage*. Cambridge: Cambridge University Press.

Bindi, R., Calzolari, N., Monachini, M., Pirrelli, V. and Zampolli, A. 1994. Corpora and computational Lexica; integration of different methodologies of lexical knowledge acquisition. *Literary and Linguistic Computing* 9 (1):29-46.

Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2004. *Structural Properties of the African Web* 2004 [cited 27 April 2004 2004]. Available from http://www2002.org/CDROM/poster/164/.

Borin, L. 2000. A corpus of written Finnish Romani texts. Paper read at Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, 2000, Athens, Greece: ELRA.

Brew, C. and McKelvie, D. 1996. Word-pair extraction for lexicography. Paper read at NeMLaP'96. Ankara, Turkey.

Brown, T.J. 1925. *English Dictionary*. Johannesburg: Pula Press.

Burchfield, R.W. (ed.). 1987. *Studies in lexicography*. Oxford; Oxford University Press.

Burnard, L. 1995. *Users Reference Guide for the British National Corpus version 1.0.* Oxford: Oxford University Press.

Burnard, L. 2002. Where did we go wrong? A retrospective look at the British National Corpus. Paper read at the 4th International conference on teaching and language corpora by doing corpus analysis: Oxford, UK.

Bussman, H. 1996. Routledge Dictionary of language and linguistics. London Routledge.

Butler, C. 1998. Collocational Frameworks in Spanish. *International Journal of Corpus Linguistics* 3 (1):1-32.

Butler, J. 1997. Excitable Speech: A politics of the performative. Routledge: New York.

Campbell, J. 1815. *Travels in South Africa, undertaken at the request of the London Missionary Society: being a narrative of the second journey in the interior of that country*. London: Guthrie.

Cantos, P, and Sanchez, A. 2001. Lexical Constellations: What collocates fail to test. *International Journal of Corpus Linguistics* 6 (2):199-228.

Cavagliá, G. 2005. Measuring the homogeneity and similarity of language corpora. PhD thesis, Information Technology Research Institute, University of Brighton, Brighton.

Čermák, F. 1997. Czech National Corpus: A case in many contexts. *International Journal of Corpus Linguistics* 2 (2):181-197.

Čermák, F. and Křen, M. 2005. Large corpora, lexical frequencies and coverage of text. Paper read at Corpus Linguistics 2005, Birmingham, UK.

Chebanne, A. M. 2002. Glides in Setswana. *Journal of the Linguistics Association for Southern African Development Communities Universities (LASU)* 1:43-49.

Cho, S.W, and O'Grady, W. 1996. Language acquisition: the emergence of a grammar. In W. O'Grady, M. Dobrovolsky and F. Katamba (eds) *Contemporary Linguistics*, London: Longman.

Church, K, and Mercer, R. 1993. Introduction to the special issue on Computational Linguistics using Large Corpora. *Computational Linguistics* 19 (1):1-24.

Church, K. and Hanks, P. 1989. Word Association Norms, Mutual Information, and Lexicography', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*:76-83.

Clear, J. 1987. Computing. In Sinclair, J. (ed). *Looking Up: An Account of the COBUILD Project.*Glasgow: Collins ELT.

Clear, J. 1992. Corpus Sampling. In G. Leiter (ed) *New Directions in English Language Corpora*. New York: Mouton de Gruyter.

Cole, D. 1995. *Setswana-Animals and plants (Setswana-Ditshedi le ditlhare)*. Gaborone: The Botswana Society.

Cole, D.T. 1955. *An Introduction to Tswana Grammar, 9th impression*. Cape Town: Longman.

Cowie, A. P. 1999. English Dictionaries for Foreign Learners: A History. Oxford: Oxford University Press.

Créissels, J., and Chebanne, M.A. 2000. *Dictionnaire Francaise-Setswana, Thanodi Sefora-Setswana*. Mogoditshane: Tasalls Publishing.

Crowdy, S. 1991. *Spoken Corpus Design an Transcription*: Longman.

Crowdy, S. 1993. Spoken Corpus Design. *Literary and Linguistic Computing* 8:259-265.

Crowdy, S. 1994. Spoken Corpus Transcription. *Literary and Linguistic Computing* 9 (1):25-28.

Culpeper, J. 2002. Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. Paper read at Conversations in life and literature ASLA Symposium, at Universitetstryckeriet.

Dash, N.S. and Chaudhuri, B.B. 2000. The process of designing a multidisciplinary monolingual sample corpus, *International Journal of Corpus Linguistics* Vol. 5 (2):179-197.

De Beaugrande, R. 1997. Text linguistics, discourse analysis, and the discourse of dictionaries. *Bibliotheque des cahiers de l'institut de linguistique de Louvain* 87:57-75.

De Haan, P. 1992. The optium corpus sample size. In *Topics in English Linguistics*, edited by G. Leitner. Berlin and New York: Mouton de Gruyter.

De Monnink, I. 1999. Combining corpus with experimental Data. *International Journal of Corpus Linguistics* 4 (1):77-111.

DeRose, S.J. 1991. "An analysis of probabilistic grammatical tagging methods." In *English Computer Corpora: selected papers and research guide*, Stig Johansson and Anna-Brita Stenström (eds). New York: Mouton de Gruyter, pp. 9- 14. Volume 3 of *Topics in English Linguistics*, ed. by Jan Svartvik and Herman Wekker.

De Schryver, G-M. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic journal of African Studies* 11 (2):266-286.

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

De Schryver, G-M, and Prinsloo, D.J. 2000a. 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. and Prinsloo D.J. 2000b. *Creating Electronic corpora for the South African Languages*. Available at: http://www.up.ac.za/academic/libarts/afrilang/generalcorpora.htm .

De Schryver, G-M, and Prinsloo, D.J. 2001. Taking Dictionaries for Bantu Languages into the New Millennium – with special reference to Kiswahili, Sepedi and isiZulu in J.S. Mdee & H.J.M. Mwansoko (eds.). 2001. *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam:188–215.

De Schryver, G-M, and Joffe, D. 2004. On how electronic dictionaries are really used. Proceedings of the 11th EURALEX conference on Lexicography, Université de Bretagne Sud, Lorient, France, 6-10 July 2004

Demuth, K, and Johnson, M. 1989. Interactions between discourse functions and agreement in Setawana. *Journal of African languages and linguistics* 11:22-35.

Dent, G. R. 1992. *Compact Setswana Dictionary: English-Setswana, Setswana-English*. 1st ed. Pietermaritzburg: Shutter & Shooter.

Diaz de Ilarraza, A. D, Gurrutxaga, A., Hernaez, A., Lopez de Gerenu, N., and Sarasola, K. 2003. HIZKING21: Integrating language engineering resources and toools into systems with linguistic capabilities. *TALN*:11-14.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1):61-74.

Evert, S. 2004. The Statistics of word coorccurrences: word pairs and collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart.

Evert, S. and Baroni, M. 2005. Testing the extrapolation quality of word frequency models. Paper read at The Corpus Linguistics Conference Series, Burmingham, UK.

Fillmore, C., Ide, N., Jurafsky, D., and Macleod, C. 1998. An American National Corpus: A Proposal. *Proceedings of the First International Language Resources and Evaluation Conference,* Granada, Spain:965-70.

Finch, G. 2000. Linguistic terms and concepts. Basingstoke. Macmillan Press.

Fletcher, W.H. 2002. Making the web more useful as a source for linguistic corpora. Paper read at Corpus linguistics in North America: Selections from Fourth North American Symposium of the American Association for Applied Corpus Linguistics.

Fletcher, W.H. 2005. Concordancing the web: promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus linguistics and the web*. Amsterdam: Rodopi.

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.

Francis, W.N, and Kucera, H. (1964, 1971, 1979). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers*. Providence: Department of Linguistics, Brown University.

Francis, W. N., Kucera, H. and Mackie, A.W. 1982. *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Fredoux, J. 1864. *A sketch of the Sechuana Grammar*. Cape Town.

Fries, U., Tottie, G. and Schneider, P. 1993. The 14th International conference on English language research on computerized corpora. Paper read at Creating and using English Language Corpora, Zurich.

Gaotlhobogwe, M. 2007. *Reteng Refutes Ramsay.* Mmegi Newspaper Online. Available

from http://www.mmegi.bw/2006/May/Wednesday10/75487749879.html.

Garside, R., Leech, G. and Sampson, G. (eds). 1987. *The Computational Analysis of English: A corpus-based approach*. Essex: Longman.

Ghani, R, Jones, R., and Mladenic, D. 2001. Mining the web to create minority language corpora. *CIKM*:279-286.

Ghani, R., Jones, R and Mladenic, D. 2001a. Automatic web search query generation to create minority language corpora. *SIGIR'01*.

Gomez, P.C. 2002. Do we need statistics when we have linguistics? *D.E.L.T.A.* 18 (2):233-271.

Gouws, R.H, and Prinsloo, D.J. 1997. Lemmatisation of Adjectives in Sepedi. *Lexikos* 7:45-57.

Gouws, R.H, and Prinsloo, D.J. 2005. Principles and practices of South African lexicography. *African Sun Media*.

Grefenstette, G. and Tapanainen, P. 1994. What is a word, What is a sentence? Problems of Tokenization. Paper read at Computational Lexicography (COMPLEX '94).

Grimes, B.F. 2004. ETHNOLOGUE: Languages of the World Fourteenth Edition. Available at: *http://www.sil.org/ethnologue/*.

Guenthner, F, and Blanco, X. (in press). Multi-Lexemic Expressions: An Overview. *Lingvisticae Investigationes Supplementa*.

Hanks, P. 1994. Linguistic Norms and Pragmatic Exploitations or, Why Lexicographers need prototype Theory, and Vice Versa. Papers in Computational Lexicography COMPLEX '94, Linguistics Hungarian Academy of Sciences, Budapest.

Hasselbring, S. 2000. *A sociolinguistics survey of the languages of Botswana*. Vol. 1, *Sociolinguistic studies of Botswana language series, Basarwa languages project*. Mogoditshane: Tassals Publishing and Books.

Hasselbring, S., Segatlhe, T and Munch, J. 2001. *A sociolinguistics survey of the languages of Botswana*. Vol. 2, *Sociolinguistic studies of Botswana language series, Basarwa languages project*. Mogoditshane: Tassals Publishing and Books.

Haztivassiloglou, V. 1994. Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistric cues to statistical word grouping system. In Hundt, M., Nesselhauf, N. and Biewer, C (eds). *The Balancing Art. Combining symbolic and statistical approaches to languages*. Cambridge: The MIT Press.

Heid, U. 1994. Relating lexicon and corpus: computational support for corpus-based lexicon building in DELIS. Paper read at EURALEX, Amsterdam, the Netherlands.

Hinton, P.R. 2004. *Statistics Explained* (2nd ed.). Essex: Routledge.

Hofland, K. and Johansson, S. (1982). Word frequencies in British and American English. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Hornby, A.S. 1996. Oxford Advanced Learner's Dictionary of Current English, 4th impression. Oxford: Oxford University Press.

Horvath, J. 1999. *Advanced Writing in English as a Foreign Language, A corpus-based study of processes and products.* [cited 20 May 2006. Available from http://www.geocities.com/writing_site/thesis/.

Hubert, P., and Labbe, D. 1988. A Model of Vocabulary Partition. *Literary and Linguistic Computing* 3 (4):223-225.

Hull, D. M. 1987. Educational development in Botswana: a plural heritage. *Journal of Negro Education* 56 (3):381-389.

Humphries, C. (ed). 2003. *Philip's Encyclopedia: Comprehensive Edition 2004*. London: Philip's.

Ide, N, and Macleod, C. 2001. The American National Corpus: A Standardized Resource for American English. Paper read at the Corpus Linguistics 2001 conference, at Lancaster University, Lancaster (UK).

Ide, N., Reppen, R, and Suderman, K. 2002. The American National Corpus: more than the web can provide. Paper read at the third Language Resources and Evaluation Conference (LREC), Las Palms, Spain.

Jespersen, O. 1909-49. A modern English grammar on historical principles, I-VII. Copenhagen. Munksgaard.

Johansson, S. 1991. Times change, and so do corpora. In Aijmer, K and Altenburg, B *English corpus linguistics: studies in honour of Jan Svartvikr*. London: Longman.

Johansson, S., and Hofland, K. 1989. *Frequency analysis of English vocabulary and grammar: based on the LOB corpus*. Oxford: Oxford University Press.

Johansson, S., and Stenstrom, A.B. (eds). 1991. *English Computer Corpora: Selected papers and research guide*. In Svartvik, J., and Wekker H. (Eds). *Topics in English Linguistics*. Vol. 3, New York: Mouton de Gruyter.

Johnson, S. 1963. *Johnson's Dictionary: A Modern Selection by E.L. McAdam, Jr. & George Milne*. London: Victor Gollancz.

Jones, D, and Plaatjie, S. 1916/1928. *The Setswana reader the tones of Setswana nouns*. Hants: Gregg International Publishers.

Jones, R, and Ghani, R. 2000. Automatically building a corpus for a minority language from the web. Paper read at 38th Meeting of the ACL, Proceedings of the Student

Research Workshop, Hong Kong.

Keller, F., Lapata, M. and Ourioupina, O. 2002. Using the web to overcome data sparseness. Paper read at the conference on empirical methods in natural language processing (EMNLP), Philadelphia.

Kennedy, G.D. 1998. *An introduction to corpus linguistics*. London; New York: Longman.

Kgasa, M.L.A. 1976. *Thanodi ya Setswana ya Dikole*. Cape Town: Longman.

Kgasa, M.L.A., and Tsonope, J. 1998. *Thanodi ya Setswana*. Gaborone: Longman.

Kilgarriff, A. 1996. Which words are particular of a text? A survey of statistical approaches. Proceedings. AISB Workshop on Language Engineering for Document Analysis and Recognition, 1996, at Sussex University:33-40

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

Kilgarriff, A. 1997a. Using word frequency lists to measure corpus homogeneity and similarity between corpora. Paper read at ACL SIGDAT workshop on very large corpora, Beijing and Hong Kong.

Kilgarriff, A. 2000. Business Models for Dictionaries and NLP. *International Journal of Lexicography* 13 (2):107-118.

Kilgarriff, A. 2001. Comparing Corpora. *International Journal of Corpus Linguistics* 6 (1):97-133.

Kilgarriff, A. 2001. Web as corpus. Paper read at Corpus Linguistics 2001, at Lancaster University (UK).

Kilgarriff, A. 2001b. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. Paper read at Collocation Workshop ACL2001, Toulouse, France.

Kilgarriff, A. 2003. What Computers can and cannot do for lexicography, or Us precision, them recall. Paper read at Asian Association for Lexicography 2003, Urayasu, Chiba, Japan.

Kilgarriff, A., and Grefenstette, G. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3):333-347.

Kilgarriff, A, and Rundell, M. 2001. Lexical Profiling Software and its Lexicographic Applications: A Case Study. Paper read at 10th EURALEX International Congress, Copenhagen, Denmark.

Kilgarriff, A, and Rundell, M. 2002. Lexical profiling software and its lexicographic applications: a case study. Paper read at EURALEX 2002, Copenhagen.

Kilgarriff, A, Rychly, Smrz, P., and D Tugwell. 2004. The Sketch Engine. Paper read at EURALEX 2004, Lorient, France. Available online at: http://www.lexmasterclass.com/people/akcv.htm

Kilgarriff, A, and Salkie, R. 1996. Corpus similarity and homogeneity via word frequency. Paper read at EURALEX '96, at Gothenberg, Sweden.

Kittredge, R. 1982. Variation and Homogeneity of Sublanguages. In Kittredge, R. and Lehrberger, J. (eds) *Sublanguage: Studies of Language in Restricted Semantic Domains*, New York: Walter de Gruyter.

Kjellmer, G. 1994. Lexical differentiators of style: Experiments in lexical variability. Paper read at The Fourteenth International Conference on English Language Research on Computerised Corpora, Zurich.

Kjellmer, G. 1994a. A dictionary of English Collocations, Oxford: Clarendon Press.

Knowles, G., and Don, M.Z. 2004. The notion of a "lemma": Headwords, roots and lexical sets. *International of Journal of Corpus Linguistics* 9 (1):68-81.

Kovarik, J. 2000. How should a large corpus be built? A comparative study of closure in annotated newspaper corpora from two Chinese sources, towards building a larger representative corpus merged from representative sublanguage collections. Paper read at ACL, Hong Kong.

Kucera, H. and Nelson, W. 1965. Computational analysis of present-day American English, Brown University Press

Landau, S.I. 1984. *Dictionaries: the art and craft of lexicography*. New York: Scribner.

Landau, S.I. 1989. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.

Larson, R, and Farber, B. 2006. *Elementary Statistics: Picturing the world*. New Jersey: Pearson Prentice Hall.

Laurén, U. 2002. Some lexical features of immersion pupils' oral and written amd written narration. *Working Papers, Department of Linguistics, Lund University* 50:63-78.

Le Bac, V., Bigi, B., Besacier, L. and Castelli, C.E. 2003. Using the Web for a Fast Language Model Construction in Minority Languages. Paper read at Eurospeech 2003, Geneva.

Lee, D. 2001. Genres, Registers, Text types, Domains, and Styles: Clarifying the Concepts and Navigating a path through the BNC Jungle. *Language Learning and Technology* 5 (3):37-72.

Lee, D. (forthcoming). Computer Corpus-based Linguistics & the uninitiated postgraduate. *to appear in proceedings of the BAAL/CUP seminar: Postgraduate*

research in Applied Linguistics: The Insider Perspective.

Leech, G., Deuchar, M., Hoogenraad, R. 1982. English Grammar for today: a new introduction. Basingstoke. Macmillan Press.

Leech, G. 1991. The state of the art in corpus linguistics. In *English Corpus Linguistics*, edited by K. Aijmer and B. Altenberg. London: Longman.

Leech, G., Rayson, P. and Wilson, A. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Leitner, G. 1992. International Corpus of English: Corpus design - problems and suggested solutions. In *New Directions in English Language Corpora*, edited by G. Leitner. Berling and New York: Mouton de Gruyter.

Leon, J. 2005. Claimed and Unclaimed Sources of Corpus Linguistics. *Henry Sweet Society Bulletin* (44):36-50.

Levin, B., Song, G. and Atkins, S. 1997. Making sense of corpus data: A case study of verbs of sound. *International Journal of Corpus Linguistics* 2 (1):23-64.

Lichtenberk, F. 2003. To list or not to list: Writing a dictionary of a language undergoing rapid and extensive lexical changes. *International Journal of Lexicography* 16 (4):386-401.

Lichtenstein, H. 1928-30. *Travels in Southern Africa in the years 1803, 1804, 1805 and 1806 (a reprint of the translation from the original German by Anne Plumptre)*. Translated by Plumptre, A. Cape Town: The Van Riebeeck Society.

Lichtenstein, H. 1973. *Foundation of the Cape: [and] About the Bechuanas*: A. A. Balkema.

Livingstone, D. 1875-85. *Missionary Travels and Researches in South Africa*. London.

Louw, J.P., and Nida, E.A. 1989. *Greek-English lexicon of the New Testament: based on semantic domains*. New York: United Bible Societies.

Lynch, J (Ed). 2004. *Samuel Johnson's dictionary*. Florida: Levenger Press.

Macleod, C. and Grishman, R. 2000. The influence of corpora on lexicons: corpora use in the creation of COMLEX syntax and NOMLEX. Paper read at 9th EURALEX International Confererence, EURALEX 2000, Universität Stuttgart, Stutthart.

Macleod, C, R Grishman, and A Meyers. 2000 *Dictionaries and balanced corpora: the interdependence of resources* [cited September 2006]. Available from www.nlp.cs.nyu.edu/publication/papers/balanced.ps.

Mair, C. 1992. Problems in the compilation of a corpus of standard Caribbean English: A pilot study. In Leitner, G. (ed). *New Directions in English Language Corpora*. Berlin and New York: Mouton de Gruyter.

Malvern, R, and Richards, B. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19 (1):85-104.

Martin, J.B., and Mauldin, M. 1997. Practical and ethical issues in lexicography: Examples from the Creek dictionary project. In Pye, C. (ed) *1996 Mid-America Linguistics Conference Papers*:565-573.

Mathangwane, J.T. 2002. Reduplicatives and their tonology in Ikalanga. *Journal of the Linguistics Association for Southern African Development Communities Universities (LASU)* 1:50-61.

Matumo, Z.I. 1993. *Setswana English Setswana Dictionary*. Gaborone: Macmillan.

McArthur, T. 1986. *Worlds of reference: lexicography, learning, and language from the clay tablet to the computer*. Cambridge; New York: Cambridge University Press.

McEnery, Tony, Paul Baker, Lou Burnard, and Mark Sebba. 1999. *Mille Working Paper 4: A Comparison of LIDES and TEI Encoding Systems for mark-up of multilingual spoken language data* [cited 20/4/2004]. Available from http://bowland-files.lancs.ac.uk/monkey/ihe/mille/wp4.htm.

McEnery, T, and Wilson, A. 1996. *Corpus linguistics*, *Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T, and Xiao, Z. 2003. Fuck revisited. Paper read at Corpus Linguistics 2003, University of Lancaster (UK).

McKee, G., Malvern, D., and Richards, B. 2000. Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Lingusitic Computing* 15 (3):323-337.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Ministry of Education. 1981. *Setswana Standard Orthography of 1981*. Gaborone: Ministry of Education.

Moe, R. 2001. Lexicography and mass production. *Notes on Linguistics* 4 (3):150-156.

Moe, R. 2003. Compiling dictionaries using semantic domains. *Lexikos* 13:215-223.

Moffat, R. 1826. A Bechuana catechism, with translation of the 3rd chapter of the Gospel by John, the Lord's Prayer and other passages of scripture in that language. London.

Moffat, R. 1842. *Missionary Labours and Scenes in Southern Africa*. London: J. Snow.

Molosiwa, A. 2004. Language and literacy issues in Botswana. Michigan: Michigan State University.

Monachini, M, and Picchi, E. 1992. Tagged Corpora: A query system. Paper read at 2nd International conference on computational lexicography, COMPLEX '92, Budapest, Hungary.

Mooko, T. 2004. An investigation into the use of Setswana to teach primary school Mathematics. *Language, Culture and Curriculum* 17 (3):181-195.

Moon, R. 2007. Sinclair, lexicography, and the Cobuild project. *International Journal of Corpus Linguistics* 12:2:159-181.

Murdock, G.P. 1987. *Outline of cultural materials*. 5th ed. New Haven: Relations Area Files.

Naden, T. 1993. From Wordlist to Comparative Lexicography. *Lexikos* 3:167-190.

Nelson, G. 2004. *International Corpus of English: Markup Manual for Written texts* 2002 [cited 04 May, 2004]. Available from http://www.ucl.ac.uk/english-usage/ice/written.pdf.

Nevegina, S.B. 1998. Some problems of borrowing in the Russian language. Вестник Омского университета. 1: 72-75.

Nyati-Ramahobo, L. 1999. *The National language: A resource or a problem - The implementation of the language policy of Botswana*. Gaborone: Pula Press.

Oakes, M.P. 1998. *Statistics for corpus linguistics*, *Edinburgh textbooks in empirical linguistics*. Edinburgh: Edinburgh University Press.

Onibere, E. Morgan, A.S., Busang, E.M. and Mpoeleng, D. 2001. Human-computer interfaces design isues for a multi-lingual English speaking country - Botswana. *Interacting with Computers* 13:497-512.

Ooi, V.B.Y. 1998. *Computer corpus lexicography*, *Edinburgh textbooks in empirical linguistics.* Edinburgh: Edinburgh University Press.

Oostdijk, N. 1988. A corpus linguistics approach to linguistic variation. *Literary and linguistic Computing* 3 (1):12-25.

Otlogetswe, T. 2004. The BNC Design as a Model for a Setswana Language Corpus. *Proceedings of CLUK'04, Birmingham, UK*:193-198.

Otlogetswe, T. 2006. Challenges to issues of balanace and representativeness in African lexicography. *Lexikos* 16:145-160.

Pagano, R. 2001. *Understanding Statistics in the behavioral sciences* (6th ed.). Belmont: Wadworth, Thomson Learning.

Paikeday, T.M. 1992. O Corpora! *Lexicographica*:307-317.

Pearsall, J. 1998. *The New Oxford Dictionary of English*. Oxford: Oxford University

Press.

Peters, M.A. 1982. *Bibliography of the Tswana language*. Pretoria: State Library.

Poplack, S. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French Project. In Fasold, R.W. and Schriffrin, D. (eds). *Language Change and Variation*, Amsterdam, Philadelphia: John Benjamins Publishing Company.

Pravec, N.A. 2002. Survey of learner corpora. *ICAME Journal* 26:81-114.

Prinsloo, D.J. 2004. Revising Matumo's Setswana-English-Setswana Dictionary. *Lexikos* 14:158-172.

Prinsloo, D.J. and De Schryver, G-M. 1999. The lemmatization of nouns in African languages with special reference to Sepedi and Ciluba. *South African Journal African Languages* 19 (4):258-275.

Prinsloo, D.J., and De Schryver, G-M. 2001. 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.

Prinsloo, D.J., and Gouws, R.H. 1996. Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho. *South African Journal of African Languages* 16 (3):100-107.

Ramsay, J. 2006. *Botswana's ethnic breakdown – OP*. Online Mmegi Newspaper [cited 22 August 2006. Available at http://www.mmegi.bw/2006/May/Tuesday9/754876171530.html.

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling. In proceedings of the *workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000, Hong Kong:1-6.

Rayson, P, Berridge, D. and Francis, B. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. Paper read at 7th International Conference on Statistical analysis of textual data, at Louvain-la-Neuve, Belgium:926-936.

Rayson, P, Leech, G. and Hodges, M. 1997. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2 (1):133-152.

Rayson, P., Wilson, A. and Leech, G. 2002. Grammatical word class variation with the British National Corpus Sampler. Paper read at New Frontiers of Corpus Research: 21st International Conference on English Language Research on Computerized Corpora, Sydney.

Rayson, P. 2003. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis. Computer Science, Computing Department,

Lancaster University, Lancaster.

Reményi, A.A. 2001. Use logbooks and find the original meaning of "representativeness". Paper read at The Corpus Linguistics 2001, at UCREL, Lancaster.

Renouf, A. 1987. Corpus Development. In J. Sinclair (ed) *Looking Up: An account of the COBUILD*, London, Glasgow: COBUILD.

Republic of Botswana. 1994. *Government Paper No. 2*: *Revised National Policy on Education* Gaborone, Botswana: Government Printer.

Republic of Botswana. 2001. Central Statistics Office. 2001 census. http://www.cso.gov.bw/

Roberts, G.F. 1998. The home page as genre: a narrative approach. Paper read at Annual Hawaii International Conference on System Sciences, at Hawaii.

Roget, P.M. 1958. *Roget's Thesaurus*. Harmondsworth: Middlesex: Penguin Books.

Rundell, M. 1996. *The corpus of the future and the future of the corpus* Talk at niversity of Exeter, special conference on 'New Trends in Reference Science', [cited 17 July 2004. Available from http://www.ruf.rice.edu/~barlow/futcrp.html.

Salt, H. 1814. *A voyage to Abyssinia and travels into the interior of that country*. London.

Sanchez, A., and Cantos, P. 1997. Predictability of word forms (types) and lemmas in linguistic corpora. A case study on the analysis of the CUMBRE corpus: An 8-million-word corpus of contemporary Spanish. *International Journal of Corpus Linguistics* 2 (2):259-280.

Sandilands, A. 1953. *Introduction to Tswana*. Tigerkloof: The London Missionary Society.

Santini, M. 2003. Identifying genres on the web. Departmental Report, ITRI-03-06, Brighton: University of Brighton.

Sardinha, B. 1999. Using key words in text analysis: practical aspects. DIRECT Papers 42:1-9

Sardinha, B. 1999a. Beginning Portuguese corpus linguistics: exploring a corpus to teach Portuguese as a foreign language. *D.E.L.T.A.* 15 (2):289-299.

Sardinha, T.B. 2000. Comparing corpora with wordsmith keywords: Comparação de corpora com WordSmith KeyWords. the ESP, São Paulo, vol. 22, no 1:187-99

Savage, D. 1990. *Testing language attitudes and use*. In Bergman, T.G. (ed), *Survey reference manual*. Dallas: Summer Institute of Linguistics.

Scannell, K.P. 2007. "The Crubadan Project: Corpus building for under-resourced languages". *Cahiers du Central*, 5: 1-10.

Scott, M. 2004-2006. Oxford WordSmith Tools version 4. Oxford: Oxford University Press.

Selolwane, O.D. 2004. Ethnic structure, inequality and governance of the public sector in Botswana. *UNRISD Project on Ethnic Structure, Inequality and Governance of the Public Sector*.

Setati, M., Adler, J., Reed, Y. and Bapoo, A. 2002. Incomplete journeys: Code-switching and other language practices in Mathematics, Science and English language classrooms in South Africa. *Language and Education* 16 (2):128-149.

Sharroff, S. 2004. Methods and Tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., and Rayson, P. *Corpus Linguistics around the World*, Amsterdam: Rodopi.

Sharoff, S. 2006. Creating general purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S. (eds). WaCky! Working papers on the Web as Corpus. Gedit, Bologna. Also: http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf

Shepherd, M, and Watters, C. 1998. The evolution of cybergenres. Paper read at 31st Annual Hawaii international Conference on System Sciences, at Hawaii.

Sigley, R. 1997. Text Categories and Where You can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2 (2):199-237.

Simons, G. F. 1998. In search of task-centered software: building single purpose tools from multipurpose components: SIL Electronic working papers 1998-004.

Sinclair, J. 1989. Corpus Creation. In *Language, Learning and Community*, edited by Candlin and McNamara: NCELTR Macquire University.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. EAGLES. Preliminary Recommendations on Corpus Typology. Available at http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html. (Accessed 16.09.2006).

Sinclair, J. 2005. Corpus and Text - Basic Principles [cited 20 October 2005]. In *Developing Linguistic Corpora: a guide to good practice*, edited by M. Wynne. Oxford: Oxbow Books, available online from http://ahds.ac.uk/linguistic-corpora/.

Snyman, J.W., J.S. Shole, and J.C. Le Roux. 1990. *Dikisinare ya Setswana English Afrikaans Dictionary. Woordeboek.* Pretoria: Via Afrika Limited.

Southerland, R. H, and Katamba F. 1996. Language in social contexts. In W. O'Grady,

M. Dobrovolsky and F. Katamba (eds). *Contemporary Linguistics: An Introduction*. London, New York: Longman.

Sperberg-McQueen, C.M. and Burnard, L. (eds). 1994. *Guidelins for electronic text encoding and interchange*. Chicago and Oxford: Text Encoding Initiative.

Svensén, B. 1993. *Practical Lexicography*. Oxford: Oxford University Press.

Summers, D. 1995. *Longman Dictionary of Contemporary English.* Longman, Essex.

Summers, D. 1996. Computer lexicography: the importance of representativeness in relation to frequency. In Thomas, J. and Short, M. (eds). *Using corpora for language research*. London: Longman:260-266.

Summers, D. (ed). 1997. Longman Essential Activator. Harlow, Essex: Addison Wesley Longman.

Teubert, W. 2001. Corpus linguistics and lexicography. *International Journal of Linguistics* 6 (special issue):125-153.

Thekiso, E. 2001. A Sociolinguistic Analysis of Communication Processes in a Bilingual Court of Law in Gaborone, Botswana, PhD Thesis, University of Warwick, UK.

Thomson, N. 1989. How to read Articles which Depend on Statistics. *Literary and Linguistic Computing* 4 (1):6-11.

Thorndike, E. 1921. The Teacher's Word Book. New York: Teachers College.

Van Warmelo, N.J. 1931. *Kinship Terminology of the South African Bantu. Pretoria*. Pretoria: Government Printer.

Varadi, T. 2001. The linguistic relevance of corpus linguistics. Paper read at Corpus Linguistics 2001, at Lancaster University.

Verlinde, S. and Selva, T. 2001. Corpus-based versus intuition-based lexicography: defining a wordlist for a French learner's dictionary. Paper read at Corpus Linguistics 2001, at UCREL, Lancaster University UK.

Villasenor-Pineda, L., M. Montes-y-Gómez, M. Pérez-Coutino, and D. Vaufreydaz. 2003. A corpus balancing method for language model construction. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*:393-401.

Vintar, Š. 1999. A Lexical Analysis of the IJS-ELAN Slovene-English Parallel Corpus. In: Vintar, Š. (ed.) *Proceedings of the workshop Language Technologies – Multilingual Aspects*. Ljubljana: Faculty of Arts, 63-69.

Volk, M. 2002. Using the web as corpus for linguistic research. In *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, edited by R.

Pajusalu and T. Hennoste. University of Tartu: Department of General Linguistics.

Wells, R. A. 1973. *Dictionaries and the Authoritarian Tradition*. The Hague: Mouton & Co. N.V., Publishers.

Wierzbicka, A. 1985. *Lexicography and conceptual analysis*. Ann Arbor: Karoma Publishers.

Wilson, A. and Rayson, P. (1993). Automatic Content Analysis of Spoken Discourse. In: C. Souter and E. Atwell (eds), Corpus Based Computational Linguistics. Amsterdam: Rodopi:215-226.

Wisker, G. 2001. *The postgraduate research handbook*. New York: Palgrave.

Wookey, A.J. 1904. *Setswana and English phrases with short introduction to grammar and a vocabulary*. Cape Town: Townshend & Son.

Wynne, M. (ed). 2005. *Developing Linguistic Corpora; a guide to good practice*. Oxford: Oxbow books, available online from http://ahds.ac.uk/linguistic-corpora/.

Xiao, Z., and A. McEnery. 2005. Two approaches to genre analysis. *Journal of English Linguistics* 33 (1):62-82.

Youmans, G. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style* 24:584-599.

Zgusta, L. 1971. *Manual of lexicography*. [Scientific ed, *Janua linguarum. Series maior, 39*. The Hague, Mouton.

Zipf, G. K. 1949. *Human behavious and the principle of least effort*. Cambridge MA: Addison-Wesley.

Zipf, G. K. 1965. *The psycho-biology of language*. Cambridge MA: MIT Press.

Zwicky, A.M., and A.D. Zwicky. 1982. Registers as a Dimension of Linguistic Variation. In in Kittredge, R. and Lehrberger, J. (eds). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin, New York: Walter de Gruyter.

Appendix 1: Proposed subentries of *pelo* headword

1. ama pelo,
2. balabala ka pelo,
3. baya pelo,
4. beta pelo,
5. betwa ke pelo,
6. bofa pelo,
7. bolawa ke pelo,
8. bolwetse jwa pelo,
9. bona pelo,
10. bongwefela jwa pelo,
11. bonosi jwa pelo,
12. boteng jwa pelo,
13. bua ka pelo,
14. bula pelo,
15. busa pelo,
16. fela pelo,
17. feretlha pelo,
18. fetola pelo,
19. gapa pelo,
20. garoga pelo,
21. kgaoga pelo,
22. go sena letsapa le fisang pelo,
23. gonolwa ke pelo,
24. isa pelo mafisa,
25. itaya pelo,
26. itse pelo,
27. kgwaralatsa pelo,
28. lala ka pelo e rotha madi,
29. mabetwa-e-pelo,
30. masetla pelo,
31. matlhomola pelo,
32. matlhotlha-pelo,
33. nametsa pelo,
34. ngomola pelo,
35. ngona pelo,
36. nna pelo,
37. nona pelo ka mathe,
38. ntsha pelo,
39. ntsha pelo pelaelo,
40. pateletsa pelo,
41. pelo e boela mannong,
42. pelo e e botlhoko,
43. pelo e e letlapa,
44. pelo e ja serati,
45. pelo e khibidu,
46. pelo e rotha madi,
47. pelo e rutha,
48. pelo e setlhogo,
49. pelo e thata,
50. pelo khutshwane,
51. pelo namagadi,
52. pelo ntsho,
53. pelo pedi,
54. pelo pholwana e a golegwa
55. pelo potsane e a golegwa
56. pelo tshweu,
57. pelo yotlhe,
58. pelo-e-thata,
59. pelo-kgale,
60. pelo-telele,
61. pelo-tlhomogi,
62. pelo-tshetlha,
63. phatlola pelo,
64. ritibatsa pelo,
65. sephiri sa pelo,
66. sera pelo,
67. sethunya sa pelo,
68. sisa pelo,
69. sulafatsa pelo,
70. swa pelo,
71. swegaswega pelo,
72. thiba maroba a pelo,
73. thuba pelo,
74. tlala pelo,
75. tlalelana pelo,
76. tlhomola pelo,
77. tlola pelo,
78. tshwara ka pelo,
79. tshwara pelo,
80. tswa pelo,
81. tswela pelo,
82. uba pelo,
83. wa pelo,
84. wela pelo

Appendix 2: Participation consent form

UNIVERSITY OF BOTSWANA & UNIVERSITY OF BRIGHTON

Recording Speech

Thank you very much for agreeing to take part in this project. The study is being carried out by Thapelo Otlogetswe, a lecturer in the Department of English of the University of Botswana currently pursuing doctoral studies with the University of Brighton, UK. His research will go a long way in compiling a national treasury of the Setswana language which will inform Setswana dictionary writers and language researchers on how words are used in ordinary, everyday conversation. This resource will provide a record of how the Setswana language is currently spoken.

We are asking a large cross-section of people around the country to help with this task by recording their own conversations. These will then be transcribed on computer and built into a database which will contain several million words, and will be used for language research.

Confidential information like personal names, phone numbers and addresses will be deleted from the tapes and transcripts.

What we would like you to do is to record your conversations using the personal stereo provided. You will also need to write down some details of all conversations you have in the forms provided.

If you have any problems with recording or filling in the form ring Thapelo Otlogetswe on 71859452 or the Secretary of the Department of English at 355 2624 who will be able to help you.

- I agree to take part in this research which is to record my conversation with others.
- I am aware that my recorded conversation will become part of a collection of texts that will be used research.
- I understand that I am free to withdraw from the investigation at any time.

Name (please print) ……

Signed ……

Date …….

OTHER SPEAKERS ON THE TAPE SHOULD GRANT APPROVAL TO BE RECORDED BY WRITING THEIR NAMES AND SIGNING BELOW.

1. ……
2. ……
3. ……
4. ……
5. ……
6. ……

Appendix 3: Conversation log

On this page please write in details of conversations recorded on TAPE NO____ SIDE____

Date started recording on this side of the tape ___/ ___ /___ (e.g. 30th January 2003 = 30/01/03)

Time started recording on this side of the tape ___ am/pm

Conversation took place in _____ cattlepost/lands/village/city

What were you doing during the conversations? – e.g. paperwork at work, cooking at home, relaxing at home, travelling by bus etc. THIS INFORMATION IS FOR USE BY THE RESEARCHER ONLY.

WRITE BELOW WHAT YOU WERE DOING WHILE RECORDING ON THIS SIDE OF TAPE.

In the space below please write in the first names or initials and details (where you know them) of the people speaking on this side of the tape. Do not forget to include YOUR OWN details.

	FIRST NAME	OCCUPATION	AGE	SEX	TRIBE	RELATIONSHIP to yourself (wife, son, friend)
1						
2						
3						
4						

Mo tsebeng e, kwala dintlha ka kgatiso e o e dirileng mo khaseteng ya nomore ____

lotlhakore____
Kgwedi le letsatsi tse kgatiso e similotseng ka tsone __/ ___/ ___ (sekai: 30 Ferikgong 2003 = 30/01/03)
Nako e kgatiso e simolotsweng ka yone ____ am/pm

Puisanyo e e gatisitswe kwa morakeng/masimong/motse/toropo ya _____

O ne o dira eng fa o gatisa puisanyo e? (sekai: o theogetse mo ofising, o iketlile kwa gae, o le mo baseng, o tlhatswa, jalo jalo)

KWALA FA TLASE SE O NENG O SE DIRA FA O GATISA LETLHAKORE LE LA KHASETE.

Fa tlase kwala maina le tse dingwe ka ga batho ba ba buiwang mo lotlhakoreng lo lwa khasete. O seka wa lebala go kwala leina la gago.

	LEINA LA NTLHA	TIRO	DINGWAGA	BONG	MORAFE	KAMANO ya gago le babui (monnao, tsala)
1						
2						
3						
4						

Appendix 4: Headteacher's letter

The Headteacher
XXXX Primary School
P. O. Box XXX
Mochudi
Botswana

19 August 2004

Dear Sir/Madam

RE: REQUEST FOR RECORDING SPOKEN SETSWANA

I am writing to request permission to record classroom interaction in Setswana classes in your school. These interactions will be transcribed for inclusion in the creation of a Setswana Corpus which will be analysed as part of doctoral studies. A corpus is a collection of texts in a computer (or digital form) for linguistic analysis.

I am a Linguistics lecturer in the Department of English at the University of Botswana. Currently I am pursuing doctoral studies at the Information Technology Research Institute, University of Brighton, UK. My research is in CORPUS LEXICOGRAPHY – the use of huge language databases for the study and creation of dictionaries and dictionary resources.

The nature of my research requires varieties of texts running into millions of words. These texts are usually obtained from newspapers, magazines, conversations, novels, plays, speeches, radio news, classroom interactions and many other sources. To build such a database, I have so far received texts from Macmillan [about 1 million words], *Mokgosi* newspaper [over 1 million words], *Naledi* newspaper [*Mmegi* newspaper insert], Department of Information and Broadcasting, and from different departments in Botswana. I am therefore making an appeal to you, that I come and record Setswana classroom interactions in your school.

The recording of Setswana classroom interactions is part of a larger project of capturing spoken Setswana in which a large cross-section of people around the

country is taking part by recording their own conversations. These will then be transcribed on computer and built into a database which will contain several million words, and will be used for scientific research which will inform PhD research and in the long term improve dictionaries and language research.

Such a study raises some ethical issues. To take care of these issues, recordings from schools will be stored in a computer completely anonymously. No one will know who has used which words. Names of teachers, students or schools will not be entered into the computer and do not have to be recorded. There will be no association between any school and any recordings to protect the integrity and privacy of the school, especially that such an association is irrelevant to the study. What will be recorded, however, is the region from which the recording is done [e.g. Southern]; the type of school [primary or secondary], the level or class of students [e.g. std 6 or form 5] and the regional origins of the teacher [e.g. Southern, Kweneng, Central, or South-East]. All participating schools in this research will be suitably acknowledged for aiding research in the Setswana language.

These recordings and transcriptions together will provide a record of how the Setswana language is spoken currently.

Any enquiry or query related to this research should be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively, please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw.
You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE
University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I hope you will be able to grant us this permission and in so doing aid the development of an important national resource. Please find attached information outlining the logistics of executing the recording process.

Yours sincerely

_____

Appendix 5: Accompanying details for classroom recordings

This page briefly outlines the process to be taken in making the proposed recordings of Setswana language classes and how such recordings may be attempted without disrupting the smooth running of teaching and learning.

We propose and envisage the following:

1. That details of the proposed research should be discussed between Setswana teacher(s) and the School Head before a trip for recording is made to the school.
2. That the researcher establishes telephonic contact with the School Head or the delegated person to agree on a specific day to travel to the school.
3. We propose to make recordings on a day between mid-September to mid-November.
4. The researcher comes to the school and establishes contact with the School Head and the relevant teachers, or at least some of them.
5. That as much as possible all the recording in a school should be completed in a single day.
6. Teachers should record their teaching themselves.
7. The researcher does not have to go into any classroom. We hope this will reduce tension on the part of the teacher since they will not feel observed.
8. A personal recorder will be provided to the teacher to take to class to record themselves.
9. At least 5 instances of teaching in a school should be recorded.
10. It is important to emphasise that this research will in no way make judgements of a pedagogical nature. We will not make judgements of whether the content delivered is relevant or judgements on a teacher's voice projection. Such considerations fall outside the remit of this research. What is important is that the Setswana language is used. What is taught and how it is taught, is not significant to this research. It must be emphasized that what will be quantified in this study is words and their usage only.

Any enquiry or query related to this research may be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively

please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw. You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE
University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I will be happy to discuss details of this research with you or any teacher who may want to find out more.

With many thanks

_____

THAPELO J. OTLOGETSWE [MR.]

Appendix 6: Letter to publishers asking for text

Dear Sir/Madam

RE: REQUEST FOR SETSWANA WRITTEN TEXT

I am writing to request permission to access your Setswana written classes in your school. The Setswana text will be included in the creation of a Setswana corpus which will be analysed as part of doctoral studies. A corpus is a collection of texts in a computer (or digital form) for linguistic analysis. The corpus will also aid Setswana research beyond PhD research.

I am a Linguistics lecturer in the Department of English at the University of Botswana. Currently I am pursuing doctoral studies at the Information Technology Research Institute, University of Brighton, UK. My research is in CORPUS LEXICOGRAPHY – the use of huge language databases for the study and creation of dictionaries and dictionary resources.

The nature of my research requires varieties of texts running into millions of words. These texts are usually obtained from newspapers, magazines, conversations, novels, plays, speeches, radio news, classroom interactions and many other sources. To build such a database, I have so far received texts from Macmillan [about 1 million words], *Mokgosi* newspaper [over 1 million words], *Naledi* [*Mmegi* newspaper insert], Department of Information and Broadcasting, and from different departments in Botswana. I am therefore making an appeal to you for more Setswana text.

The Setswana text that we are requesting will be part of a larger project of capturing the use of varieties of Setswana

Any enquiry or query related to this research should be left with the Secretary to the Department of English, University of Botswana at: Tel: 355 2624. Alternatively, please fax any question or query to: 5985 098, or email me at thapelo.otlogetswe@itri.brighton.ac.uk or otlogets@mopipi.ub.bw.
You are also welcome to contact me directly at the following address:

THAPELO J. OTLOGETSWE

University of Botswana
Department of English
Private Bag UB 00703
Gaborone

I hope you will be able to grant us access to the text in so doing aid the development of an important national resource. Please find attached information outlining the logistics of executing the recording process.

Yours sincerely

_____

THAPELO J. OTLOGETSWE [MR.]

Appendix 7: BNC Part-of-speech codes

(from www.kilgariff.co.uk)

AJ0

adjective (general or positive) e.g. *good, old*

AJC

comparative adjective e.g. *better, older*

AJS

superlative adjective, e.g. *best, oldest*

AT0

article, e.g. *the, a, an, no* . Note the inclusion of *no*: articles are defined as determiners which typically begin a noun phrase but cannot appear as its head.

AV0

adverb (general, not sub-classified as **AVP** or **AVQ**), e.g. *often, well, longer, furthest*. Note that adverbs, unlike adjectives, are not tagged as positive, comparative, or superlative. This is because of the relative rarity of comparative or superlative forms.

AVP

adverb particle, e.g. *up, off, out*. This tag is used for all prepositional adverbs, whether or not they are used idiomatically in phrasal verbs such as "Come out here", or "I can't hold out any longer".

AVQ

wh-adverb, e.g. *when, how, why*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

CJC

coordinating conjunction, e.g. *and*, or, *but*.

CJS

subordinating conjunction, e.g. *although, when*.

CJT

the subordinating conjunction *that*, when introducing a relative clause, as in "the day that follows Christmas". Some theories treat *that* here as a relative pronoun; others as a conjunction. We have adopted the latter analysis.

CRD

cardinal numeral, e.g. *one, 3, fifty-five, 6609*.

DPS

possessive determiner form, e.g. *your, their, his*.

DT0

general determiner: a determiner which is not a **DTQ** e.g. *this* both in "This is my house" and "This house is mine". A *determiner* is defined as a word which typically occurs either as the first word in a noun phrase, or as the head of a noun phrase.

DTQ

wh-determiner, e.g. *which, what, whose, which*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

EX0

existential *there*, the word thereappearing in the constructions "there is...", "there are ...".

ITJ

interjection or other isolate, e.g. *oh, yes, mhm, wow*.

NN0

common noun, neutral for number, e.g. *aircraft, data, committee*. Singular collective nouns such as *committee* take this tag on the grounds that they can be followed by either a singular or a plural

verb.

NN1

singular common noun, e.g. *pencil, goose, time, revelation*.

NN2

plural common noun, e.g. *pencils, geese, times, revelations*.

NP0

proper noun, e.g. *London, Michael, Mars, IBM*. Note that no distinction is made for number in the case of proper nouns, since plural proper names are a comparative rarity.

ORD

ordinal numeral, e.g. *first, sixth, 77th, next, last*. No distinction is made between ordinals used in nominal and adverbial roles. *next* and *last* are included in this category, as general ordinals.

PNI

indefinite pronoun, e.g. *none, everything, one* (pronoun), *nobody*. This tag is

applied to words which always function as heads of noun phrases. Words like *some* and *these*, which can also occur before a noun head in an article-like function, are tagged as determiners, **DT0** or **AT0**.

PNP

personal pronoun, e.g. *I, you, them, ours*. Note that possessive pronouns such as *ours* and *theirs* are included in this category.

PNQ

wh-pronoun, e.g. *who, whoever, whom*. The same tag is used whether the word is used interrogatively or to introduce a relative clause.

PNX

reflexive pronoun, e.g. *myself, yourself, itself, ourselves*.

POS

the possessive or genitive marker *'s* or *'*. Note that this marker is tagged as a distinct word. For example, "Peter's or someone else's" is tagged

PRF

the preposition *of*. This word has a special tag of its own, because of its high frequency and its almost exclusively postnominal function.

PRP

preposition, other than *of*, e.g. *about, at, in, on behalf of, with*. Note that prepositional phrases like *on behalf of* or *in spite of* are treated as single words.

TO0

the infinitive marker *to*.

UNC

"unclassified" items which are not appropriately classified as items of the English lexicon. Examples include foreign (non-English) words; special typographical symbols; formulae; hesitation fillers such as *errm* in spoken language.

VBB

the present tense forms of the verb *be*, except for *is* or *'s am, are 'm, 're, be* (subjunctive or imperative), *ai (as in ain't*).

VBD

the past tense forms of the verb *be, was, were*.

VBG

-ing form of the verb *be, being*.

VBI

the infinitive form of the verb *be, be*.

VBN

the past participle form of the verb *be, been*

VBZ

the -s form of the verb *be, is, 's*.

VDB

the finite base form of the verb *do*, *do*.

VDD

the past tense form of the verb *do*, *did*.

VDG

the -ing form of the verb *do*, *doing*.

VDI

the infinitive form of the verb *do*, *do*.

VDN

the past participle form of the verb *do*, *done*.

VDZ

the -s form of the verb *do*, *does*.

VHB

the finite base form of the verb *have*, *have, 've*.

VHD

the past tense form of the verb *have, had, 'd*.

VHG

the -ing form of the verb *have, having*.

VHI

the infinitive form of the verb *have*, *have*.

VHN

the past participle form of the verb *have*, *had*.

VHZ

the -s form of the verb *have*, *has, 's*.

VM0

modal auxiliary verb, e.g. *can, could, will, 'll, 'd, wo*(as in *won't*)

VVB

the finite base form of lexical verbs, e.g. *forget, send, live, return*. This tag is used for imperatives and the present subjunctive forms, but not for the infinitive (VVI).

VVD

the past tense form of lexical verbs, e.g. *forgot, sent, lived, returned*.

VVG

the -ing form of lexical verbs, e.g. *forgetting, sending, living, returning*.

VVI

the infinitive form of lexical verbs , e.g. *forget, send, live, return*.

VVN

the past participle form of lexical verbs, e.g. *forgotten, sent, lived, returned*.

VVZ

the -s form of lexical verbs, e.g. *forgets, sends, lives, returns*.

XX0

the negative particle *not* or *n't*.

ZZ0

alphabetical symbols, e.g. *A, a, B, b, c, d*.

The following *portmanteau tags* are used to indicate where the CLAWS system has indicated an uncertainty between two possible analyses:

AJ0-AV0

adjective or adverb

AJ0-NN1

adjective or singular common noun

AJ0-VVD

adjective or past tense verb

AJ0-VVG

adjective or -ing form of the verb

AJ0-VVN

adjective or past participle

AVP-PRP

adverb particle or preposition

AVQ-CJS

wh-adverb or subordinating conjunction

CJS-PRP

subordinating conjunction or preposition

CJT-DT0

that as conjunction or determiner

CRD-PNI

one as number or pronoun

NN1-NP0

singular common noun or proper noun

NN1-VVB

singular common noun or base verb form

NN1-VVG

singular common noun or -ing form of the verb

NN2-VVZ

plural noun or -s form of lexical verb

VVD-VVN

past tense verb or past participle

The following codes are used with c elements only:

PUL

left bracket (i.e. (or [ )

PUN

any mark of separation (. ! , : ; - ? ...)

PUQ

quotation mark (` ' `` ")

PUR

right bracket (i.e.) or])

Note that some punctuation marks (notably long dashes and ellipses) are not tagged as such in the corpus, but appear simply as entity references.