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.