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

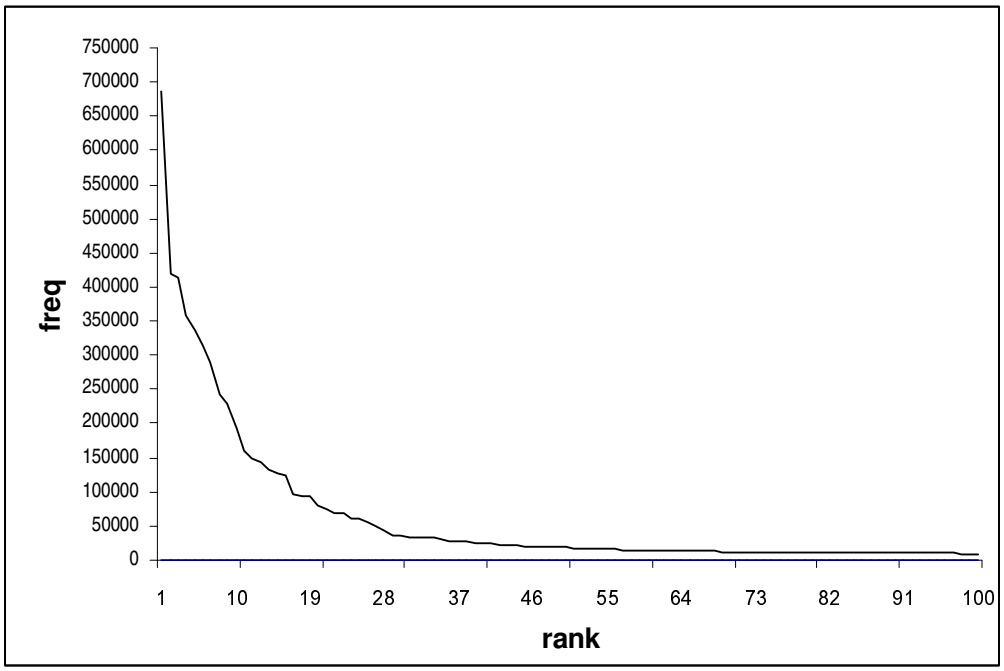
> In a Zipfian distribution, the most common item has twice as many occurrences as the second most common, three times as many as the third, a hundred times as many as the hundredth, a thousand times as many as the thousandth, and a million times as many as the millionth (Kilgarriff, 1996: 2).

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

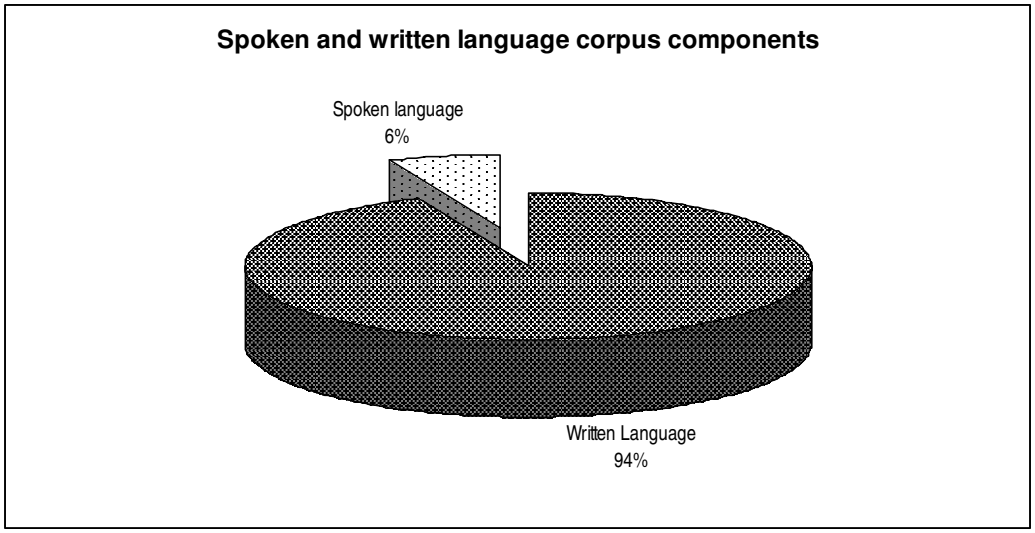
Below we give the different components of the whole corpus on the basis of tokens, types, TTR and STTR. First, we calculate the large corpus components of written and spoken language. The results are given in Table 25.

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.

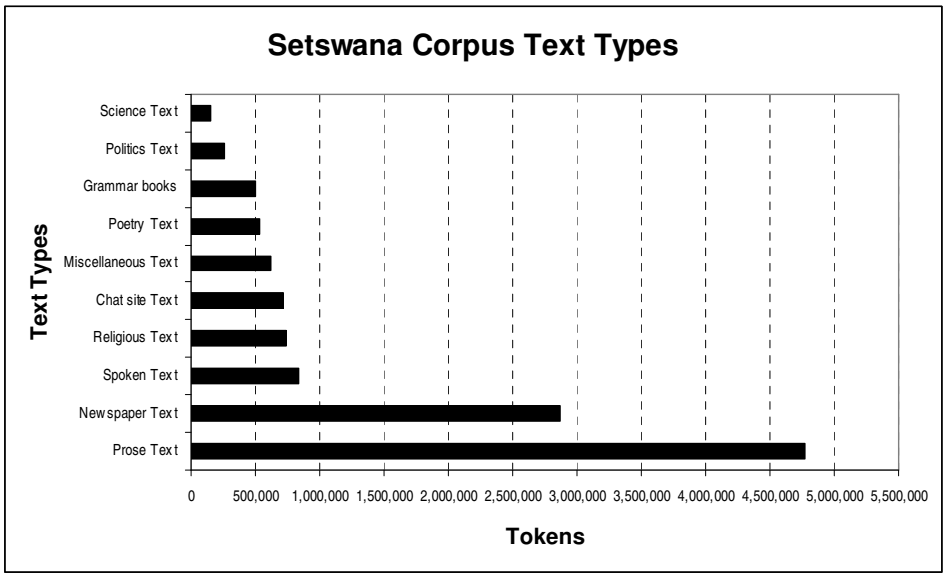
Figure 6: Spoken and written language corpus components pie chart



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

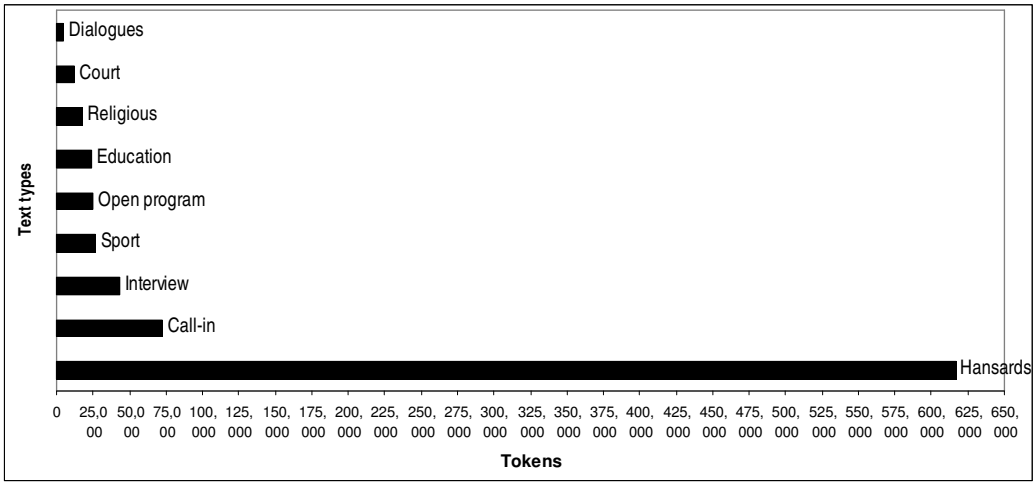
The overall spoken component of the Setswana corpus has 840,400 tokens and 38,118 types. It has a type/token ratio of 4.54 and the STTR of 32.94. Table 26 shows the breakdown parts of the spoken component.

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

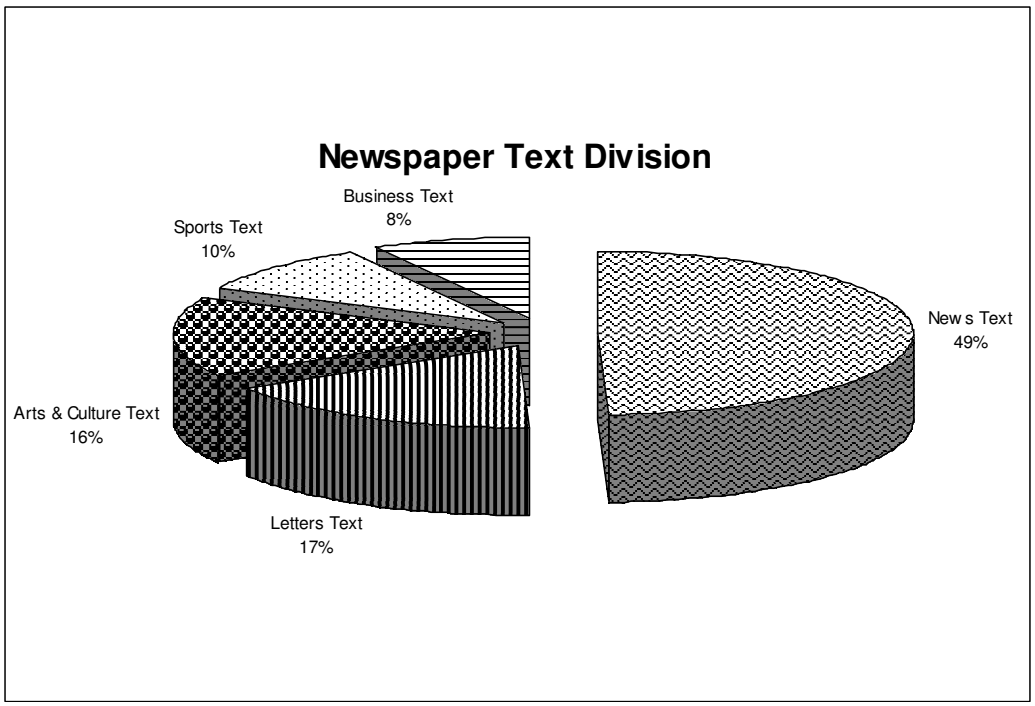


Figure 9 reveals that about 50% of the newspaper text is News while about another 30% is made up of Arts and Culture and Letters, 16% and 17% respectively. The Letters subcorpus comprises letters to the editor, editorials and newspaper columns. Business text has the smallest percentage of 8%. Business 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.