

Chapter 3

Compilation of Setswana corpora

3.1. Introduction

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

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

If African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should become an absolute priority

according to De Schryver and Prinsloo (2000a, 2000b) and Prinsloo and De Schryver (1999, 2001b). The value of corpora for the compilation of dictionaries is generally accepted in the literature:

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

3.2 Major (English) electronic corpora in historical perspective

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

3.2.1 COBUILD main corpus

(a) Aims and objectives

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

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

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

(b) Sampling principle

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

3.2.2 LOB corpus (Lancaster Oslo Bergen)

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

(a) Sampling principle

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

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

Table 1: Basic composition of Brown and LOB corpora

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

(Johansson and Hofland, 1989:2)

3.2.3 Longman Lancaster English language corpus

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

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

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

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

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

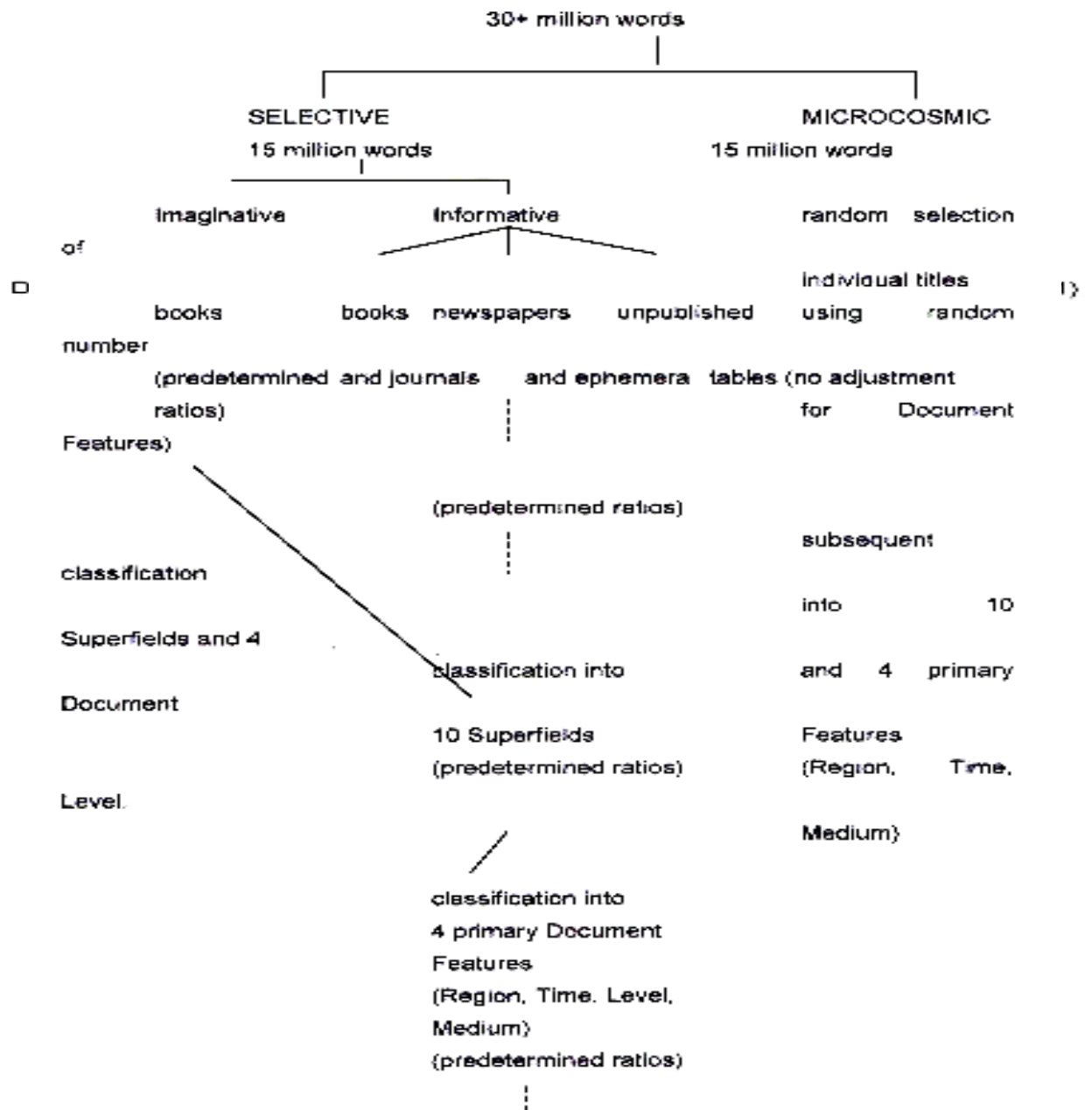


Figure 6: Longman/Lancaster English language corpus – current structure

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

3.3 Issues 'size', 'balance', 'representative' and 'organic corpora'

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

Shannon (1948:5) states:

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

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

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

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

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

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

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

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

3.3.1 Representativeness and balanced samples

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

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

3.4 Compilation of the Setswana corpus

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

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

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

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



Typical OCR output

“

ntlha ya go mo gopotsa mmaagwe ka semphato sa bontsho, a mo tsaya jaaka wa losika.

Selo se sengwe se a neng a se gakologelwa sentle ke fa a ne a mmotsa gore monnamogolo o ba tiogetse leng. Ikaterena o ne a thubega ka sona seletshogo sa makanyane a sa batle go go modiwa. O rile gore o mmolelela ka loso lwa ga rraagwe, le gore ba gaufi le go ntsha diaparo tsa gagwe kwa Tsetse, mosadimo golo a nne! a go lela a re goo! Ba ne ba eme fa makgaoganong a tsela ya kwa gamotlatla, ya Tigane le ya Ditsobotla fa kutleng ya borapharakano ba K~,holini. K~,e fa teng fa e rileng rapharaka no a ba kgalemelela go ema ba sa tshuba dipone tsa go phaka, a ba bolelela gore ba se ka ba tshwenngwa ke selelo sa gagwe. Ba mmogo ..

"A o tla bua kgotsa nnyaa?" Matlala a mo gwetlha ka potso a ithaya a re o mo file sebaka.

"A ga go na gore nka lokololwa ka beile ka tsoga ke ya gae ka tla ka boa gape? K~,ana ga go na yo o itseng kwa ke leng teng."

"Beile ya eng o ise o lebagannngwe le molato?" Matlala a mo leba a nyatsa leanonyana la phokoje a batla go ja mokoko o le mo setlhareng. "Le gale, ka moso nako e, o tla bo o o sikere ka magetla molato wa go gweba ka diritibatsi ... fa o gana go kgwa molalatlhageng."

Ga a d ka a nagana lobaka Modiko, ke fa a gotolela Matlala -matiho a mmolelela ka tshwetso ya gagwe. "Ga go na se nka se buang nao rra. l<-e ema ka gore ga ke itse sepe se ke se orisetwang kgakqamosi e e kanakana. "

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

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

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

Table 2: Typical OCR misreading of signs

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

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

3.4.1 Text encoding

Text encoding activities are forthcoming and briefly outlined here. According to Prinsloo and de Schryver (2000:13) text encoding can consist of any combination

of (a) word tokenisation, (b) part-of-speech tagging, (c) lemmatisation and (d) syntactic parsing.

(a) Word tokenisation

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

(b) Part-of-speech tagging

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

(c) Lemmatisation

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

Example 1

Locative (-ng)

Lemma

sekolo(school)

sekolong

(to school)

naga (veld)

nageng

(to the veld)

Verbs

Lemma

reka (**buy**)

rekela, rekelela, rekega, rekisa, rekisetsa.

bofa (**tie**)

bofega, bofelela a, bofolola, bofologa, bofisa

(c) Syntactic parsing

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

Example 2

<i>Monna</i>	<i>yo o</i>	<i>jang</i>
The man	who	is eating
Noun	relative concord	verb stem with relative <i>-ng</i>

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

3.4.2 Querying text corpora

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

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

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

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

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

3.4.3 Alphabetical word list and frequency word list

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

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



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

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

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

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

3.5 Compilation of a Setswana oral corpus

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

3.5.1 Keynes definition

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

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

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

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

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

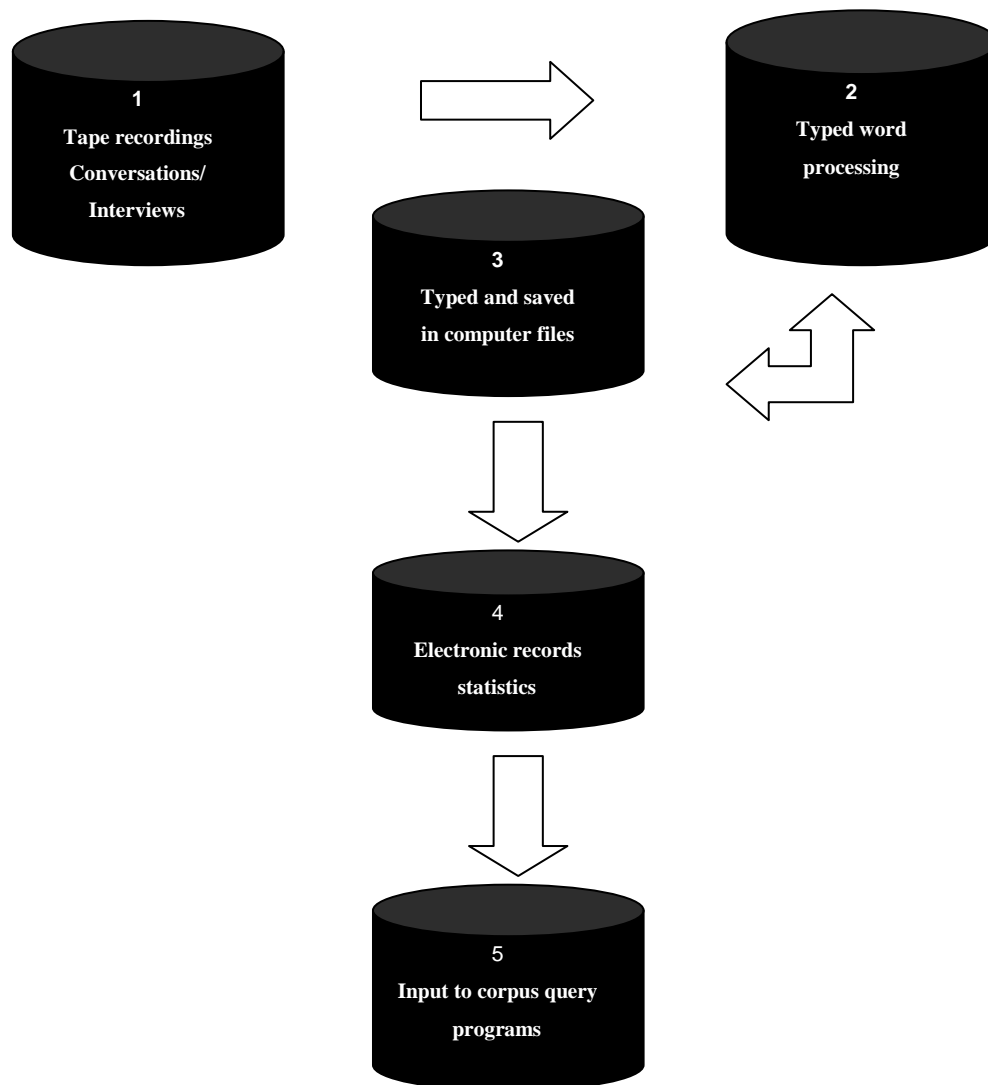


Figure 7: Flow chart representing oral corpus recordings and processing

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

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

Aneela ka bokhutswane ka tsa botshelo jwa gago. (Briefly tell about your life history)

Bapisa botshelo jwa gago le botshelo jwa segompieno. (Compare your life style with the present life style.)

Ke eng se o eletsang go bona se fedisiwa kgotsa o sa se rate? (In your own opinion, which things do you consider most irrelevant or you dislike in the modern way of life?)

A puo le setso ya Setswana di a somarelwa kgotsa di anyelela? (Do you consider the language and culture of Setswana conserved or destroyed?)

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

3.6 Summary of the Setswana oral projects

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

Project 1: An elderly lady.

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

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

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

Project 2: A male educator at a particular School.

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

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

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

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

Project 3: A certain farmer in a village.

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

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

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

Project 4: Programme director at the graduation ceremony.

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

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

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

Project 5: A lecturer at a particular university.

The narrator is a black South African who speaks other African languages and also respects other African cultures. He grew up in Botswana, taught Setswana first language at various institutions and has a very good command of the Setswana language. His lecture is based on the differences and similarities that exist between

language and culture. He is well travelled, understands various cultures and outlines the uniqueness of all traditions by their attire and use of language. The narrator singled out all family members according to their portfolio's and responsibilities. He understands all black rituals, their function and preparation according to each special occasion. He showed how different cultures cater for unrelated children and less fortunate adults to make them feel acceptable, for example, donations are given to adults to help them fulfil their obligations as parents to their children.

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

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

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

Project 6: A resident in Makapanstad area.

The narrator was born in 1938 in Lady Selborne (Pretoria). At seven years when attending grade 4, his parents decided to move to East Lynne. He was brought up by

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

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

Project 7: An educator in Makapanstad.

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

The narrator notices various Setswana language structures. The Setswana language has many dialects that have been influenced by other various cultures. Although the narrator uses the standardized form of language, she sometimes uses borrowed or loanwords. She is selfish about the Setswana language. The narrator is disturbed by the fact that the Batswana people today have to compromise their language because of technology and changes. The problem is that the Setswana language lacks scientific and modern technology terminology. According to the speaker, the Setswana spoken by the youth is not recommended since it is misleading and incorrect. She criticises

technology and encourages the Batswana people not to be depended on loanwords or borrowed words but rather to formulate their own vocabulary. The narrator further recommends that the Batswana people should go back to their roots. For example, use of traditional medicines and their names should be protected and reserved.

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

Project 8: Lesson presentation by a grade 12 educator.

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

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

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

Project 9: A traditionalist in Mabopane.

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

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

Project 10: A particular man in a village.

The narrator does not reveal where comes from nor his age. He only narrates on his brother’s life who originates from Kgwadibeng in Hammanskraal. The narrator is both

traditional and modern. The language use has no cultural background, but is infiltrated by terms that are linked to the environment in which he lives. The food that he eats is Setswana traditional and influences his typical life style. He mixes languages when addressing his experiences. His language is influenced by the environment in which he lives. He does not visualize any life changes. The language he uses is centred on events, experiences and the environment. He is a Setswana speaker by birth but influenced by the environment in which he lives. The key words *kereya* (to find) appears 11 times in the text and *mara* (but) appears 20 times. These words are loanwords and are more often used than it is expected in the oral text.

Project 11: An educator in a particular school.

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

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

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

Project 12: A lady principal at a particular school.

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

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

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

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

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

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

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

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

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

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

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



16	SE	312
17	WA	285
18	SA	275
19	KWA	235
20	MME	229
21	BO	228
22	FA	222
23	GE	220
24	TSE	214
25	NA	190
26	KO	182
27	NNA	180
28	BONA	170
29	TSA	163
30	JAANONG	147
31	LA	127
32	TENG	126
33	FELA	112
34	TLA	108
35	NTSE	99
36	YO	99
37	BANA	97
38	RA	91
39	NAKO	90
40	NENG	87
41	DIRA	86
42	KGOMO	78
43	YONA	78
44	BATHO	77
45	BUA	77
46	KGWEDI	77
47	KGOTSA	76
48	RONA	76
49	SETSWANA	75
50	MOTHO	67

66	MAINA	46
67	NGWANA	46
68	BONE	45
69	JAANA	45
70	DILO	44
71	MORAGO	44
72	FITLHELA	43
73	GAPE	43
74	LENG	43
75	MONGWE	43
76	BATLA	42
77	GAGWE	42
78	JA	42
79	DITLHAKA	41
80	GAE	41
81	GONE	40
82	MODIMO	40
83	NGWAGA	39
84	DIKGOMO	38
85	TLE	38
86	TSHWANETSE	38
87	EE	37
88	SEKOLO	37
89	TSAYA	37
90	TSENA	37
91	FITLHA	36
92	MONNA	36
93	SENA	36
94	LO	35
95	MARA	35
96	SEKOLONG	35
97	TOTA	35
98	DIANE	33
99	MALOME	33
100	MOTSWANA	33

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

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

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

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

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

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

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

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

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

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

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

Table 5: Positive keyness versus frequency in the oral corpus

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

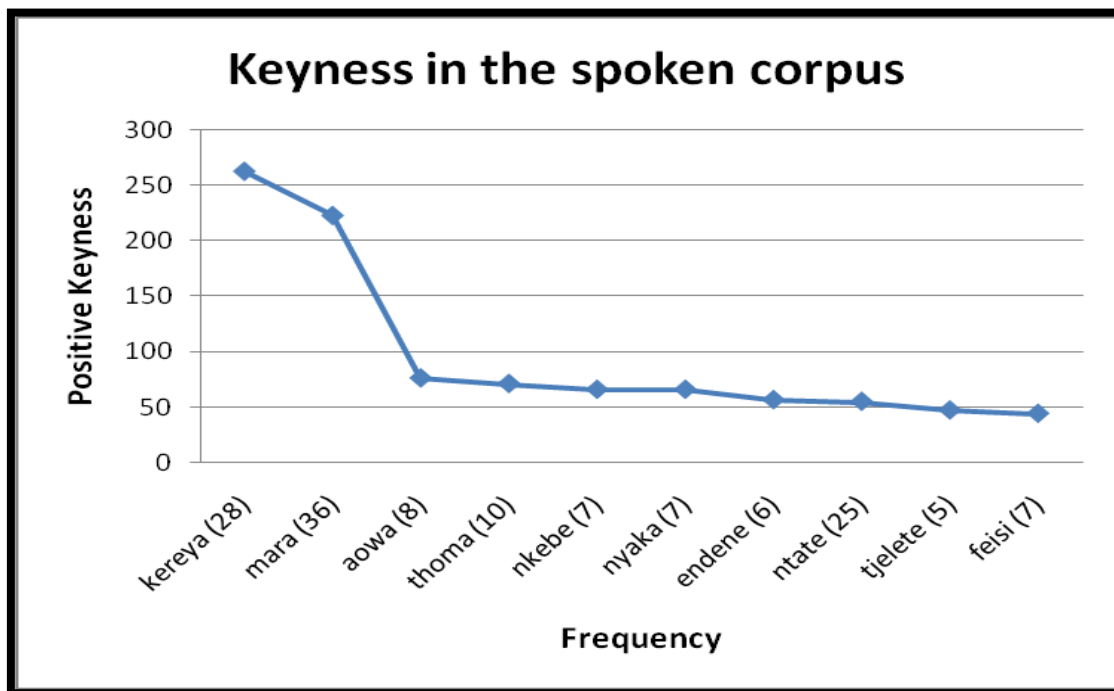


Figure 8: Graphical representation of positive keyness versus frequency in the spoken corpus

Positive keyness refers to those words used much more frequently than it is expected in a given text. The most prominent ones with the highest frequency will firstly be discussed i.e. words like *kereya* (28) and *mara* (36). These words stand out as being unusually frequent and have the highest keyness as shown in Figure 8. Taking the second example of words in the same category like *ntate* (25) and *thoma* (10), one realises the inconsistencies in terms of the likeness between frequency and keyness. One notices that the majority of these words such as *kereya* and *mara* are loanwords. For example, the words *kereya* and *mara* derive from the Afrikaans words ‘*kry*’ and ‘*maar*’ respectively while words such as *ntate* and *thoma* are influences from other dialects and other Sotho languages such as Sepedi and Southern Sotho. These loanwords have the lowest keyness. It is also important to note that these words are used when people are in a more relaxed atmosphere, for example, we often do not find words such as *bona*, *lemororo*, *nyaa* and *simolola* in their regular use in the language. It is important to state that dialects of other regions or social classes be taken into

consideration during dictionary compilation. For example, *ntate* is more frequently used than *rre* but both words refer to the word (father) where *ntate* is more often spoken by the Bafokeng in Rustenburg. Conclusion can thus be drawn that there is no clear link between the word frequency and the positive keyness in the oral corpus. The spoken words in Table 5 will now be compared with their more natural counterparts in the language. Consider Table 6 in this regard:

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

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

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

Table 7: Negative keyness versus frequency in the oral corpus

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

<i>motho</i> (person)	67	- 39.2
<i>omongwe</i> (somebody)	46	- 29.9
<i>Mopitlwe</i> (March)	6	- 24

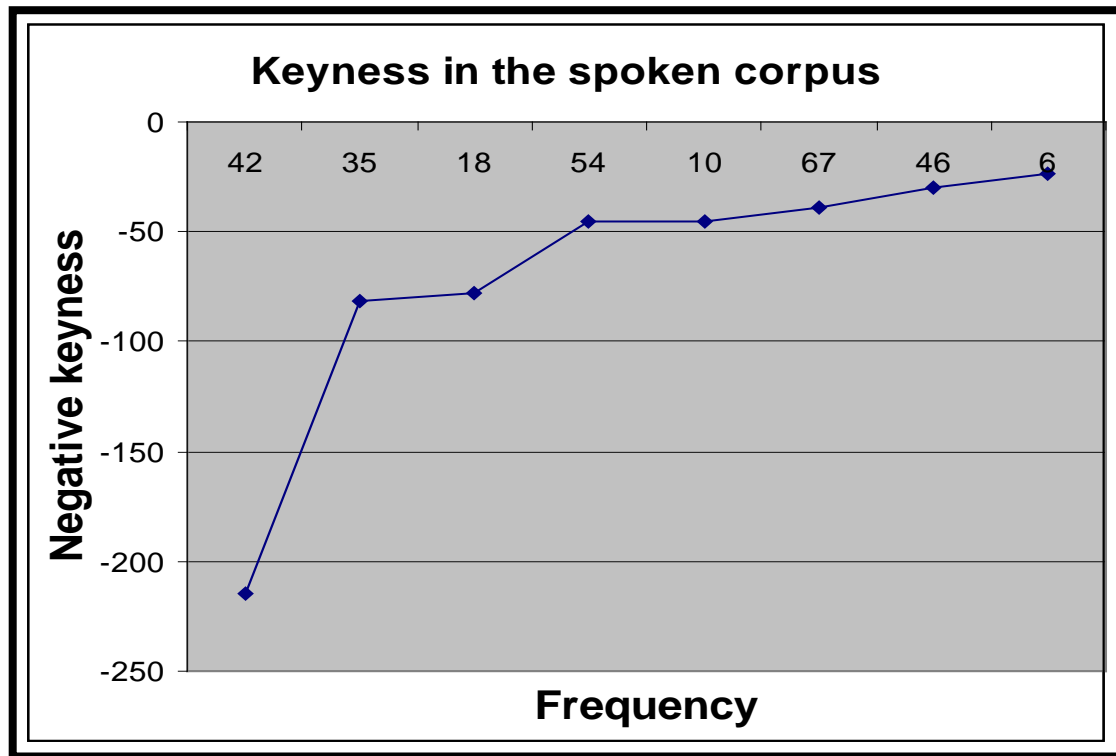


Figure 9: Graphical representation of negative keyness

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

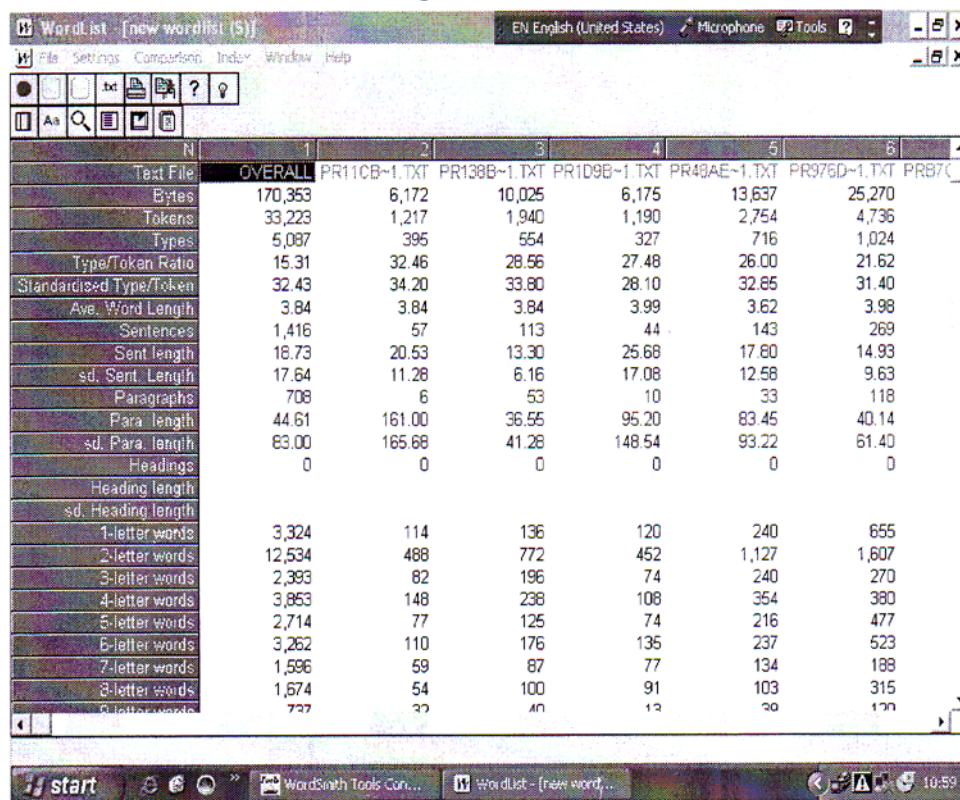
"A word which is negatively key occurs less often than would be expected by chance in comparison with the reference corpus".

One can follow the same approach in Table 7 of comparing word frequency in relation to the Keyness function, taking for example words with the highest frequencies such as *motho* (67), and *gagwe* (42), with keyness of (-39), and (-214.9) respectively. The difference between *motho* and *gagwe* in terms of their keyness does not show any correlation between keyness and frequency. If one goes further in comparing for example, words such as *Mopitlwe* (6) with the lowest frequency and *motho* (67) with

the highest frequency, it is important to state that *Mopitlwe* with the lowest frequency has the highest keyness while *motho* with the highest frequency has a relatively high keyness. Thus it is important to conclude from the above given examples that there is no correlation between the negative keyness and the word frequency in the spoken corpus.

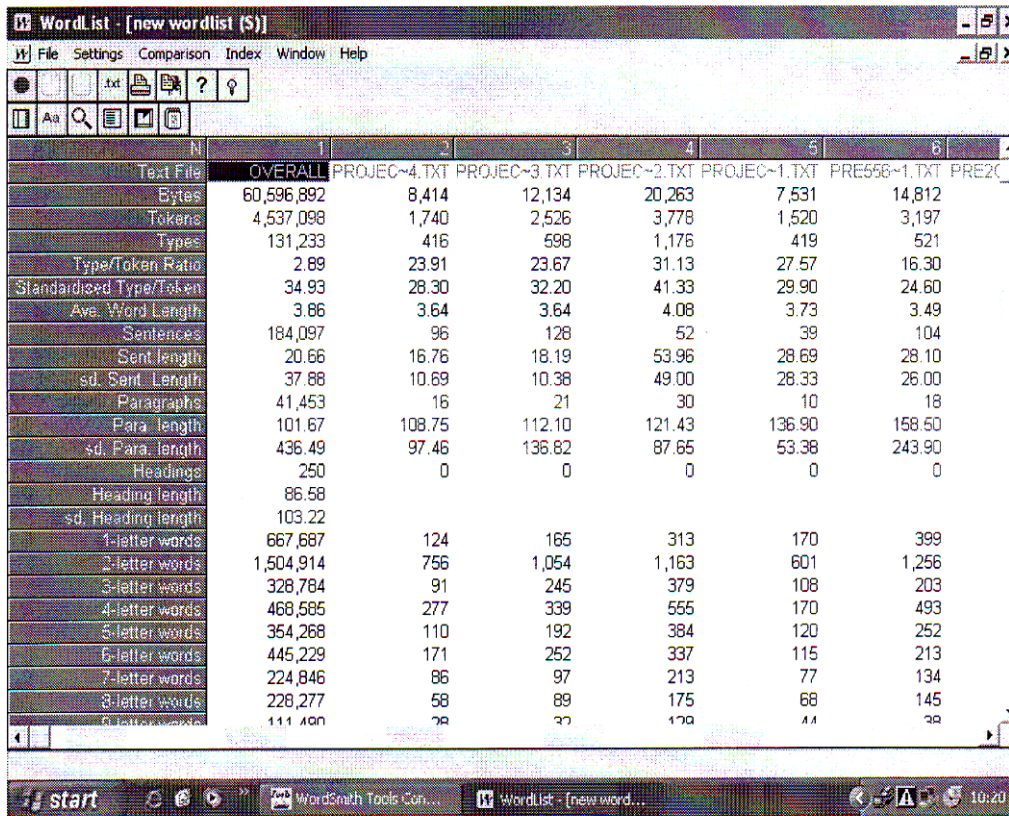
3.8 Comparison between the South African oral corpus and the South African written corpus

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



N	1	2	3	4	5	6	
Text File	OVERALL	PR11CB-1.TXT	PR138B-1.TXT	PR1D9E-1.TXT	PR48AE-1.TXT	PR976D-1.TXT	PR87C-1.TXT
Bytes	170,353	6,172	10,025	6,175	13,637	25,270	
Tokens	33,223	1,217	1,940	1,190	2,754	4,736	
Types	5,087	395	554	327	716	1,024	
Type/Token Ratio	15.31	32.46	28.56	27.48	26.00	21.62	
Standardised Type/Token	32.43	34.20	33.80	28.10	32.65	31.40	
Ave. Word Length	3.84	3.84	3.84	3.99	3.62	3.98	
Sentences	1,416	57	113	44	143	269	
Sent length	18.73	20.53	13.30	25.68	17.80	14.93	
sd. Sent. Length	17.64	11.28	6.16	17.08	12.58	9.63	
Paragraphs	708	6	53	10	33	118	
Para. length	44.61	161.00	36.55	95.20	83.45	40.14	
sd. Para. length	83.00	165.68	41.28	148.54	93.22	61.40	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	3,324	114	136	120	240	655	
2-letter words	12,534	488	772	452	1,127	1,607	
3-letter words	2,395	82	196	74	240	270	
4-letter words	3,853	148	238	108	354	380	
5-letter words	2,714	77	125	74	216	477	
6-letter words	3,262	110	176	135	237	523	
7-letter words	1,596	59	87	77	134	188	
8-letter words	1,674	54	100	91	103	315	
9-letter words	737	32	40	13	30	120	

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



Text File	OVERALL	PROJEC~4.TXT	PROJEC~3.TXT	PROJEC~2.TXT	PROJEC~1.TXT	PRE556~1.TXT	PRE20~1.TXT
Bytes	60,596,892	8,414	12,134	20,263	7,531	14,812	
Tokens	4,537,098	1,740	2,526	3,778	1,520	3,197	
Types	131,233	416	598	1,176	419	521	
Type/Token Ratio	2.89	23.91	23.67	31.13	27.57	16.30	
Standardised Type/Token	34.93	28.30	32.20	41.33	29.90	24.60	
Ave. Word Length	3.86	3.64	3.64	4.08	3.73	3.49	
Sentences	184,097	96	128	52	39	104	
Sent length	20.66	16.76	18.19	53.96	28.69	28.10	
sd. Sent Length	37.88	10.69	10.38	49.00	28.33	26.00	
Paragraphs	41,453	16	21	30	10	18	
Para length	101.67	108.75	112.10	121.43	136.90	158.50	
sd. Para length	436.49	97.46	136.82	87.65	53.38	243.90	
Headings	250	0	0	0	0	0	
Heading length	86.58						
sd. Heading length	103.22						
1-letter words	667,687	124	165	313	170	399	
2-letter words	1,504,914	756	1,054	1,163	601	1,256	
3-letter words	328,784	91	245	379	108	203	
4-letter words	468,585	277	339	555	170	493	
5-letter words	354,268	110	192	384	120	252	
6-letter words	445,229	171	252	337	115	213	
7-letter words	224,846	86	97	213	77	134	
8-letter words	228,277	58	89	175	68	145	
9-letter words	111,499	28	32	129	44	38	

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

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

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

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

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



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

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

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

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



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

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

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

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

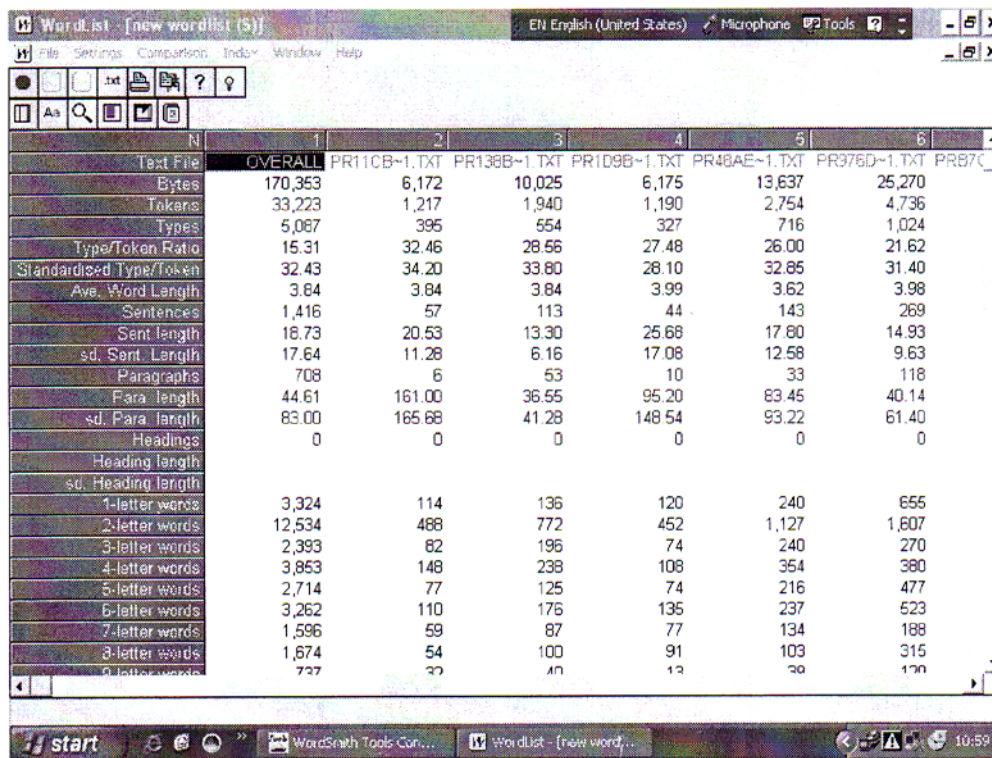
The words *pelo* and *matlho* are more formal and are more often used in the written corpus than in the spoken corpus. It is important to note that people in conversations turn to pronounce words such as *iwa* as *yiwa*. Consider the following example:

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

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

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

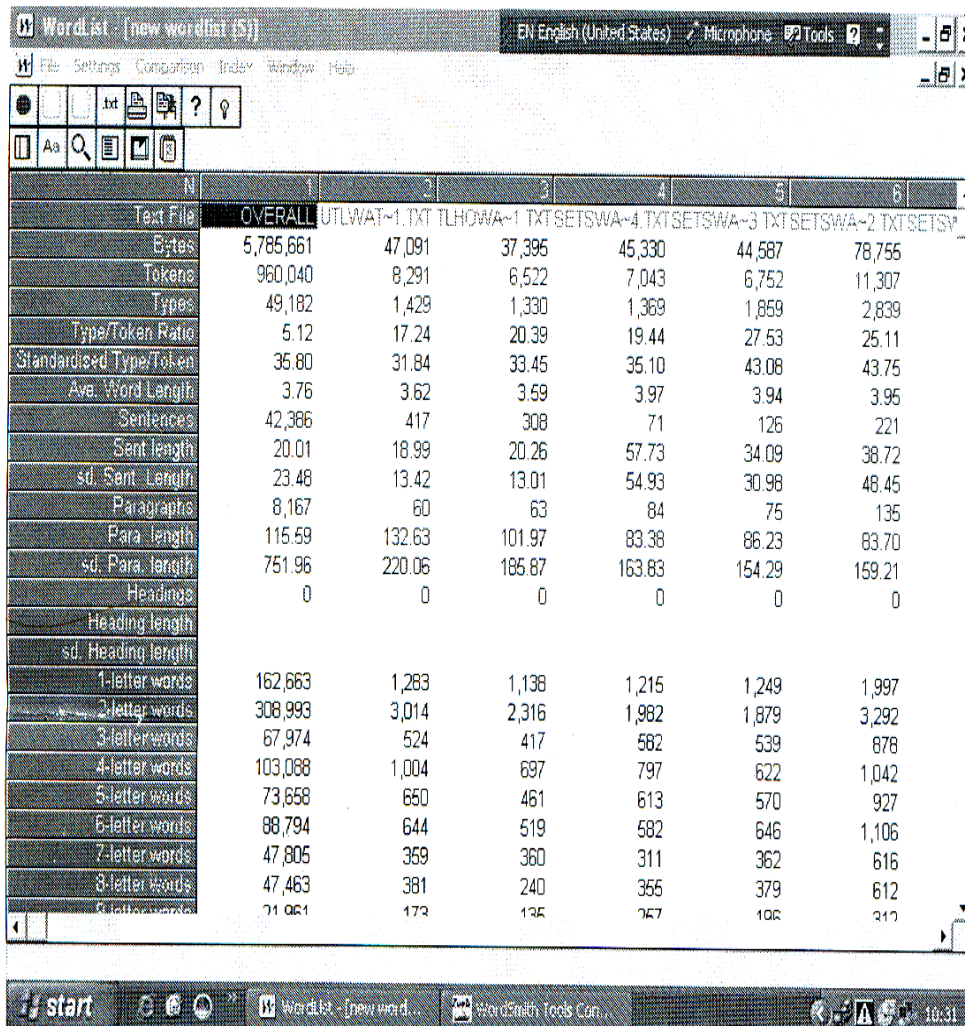
Figure 12 represents the statistical analysis of the South African oral corpus in WordSmith Tool consisting of 33,223 tokens.



N	1	2	3	4	5	6	
Text File	OVERALL	PR11CB~1.TXT	PR138B~1.TXT	PR109B~1.TXT	PR48AE~1.TXT	PR976D~1.TXT	PR97C~1.TXT
Bytes	170,353	6,172	10,025	6,175	13,637	25,270	
Tokens	33,223	1,217	1,940	1,190	2,754	4,736	
Types	5,067	395	554	327	716	1,024	
Type/Token Ratio	15.31	32.46	28.56	27.48	26.00	21.62	
Standardised Type/Token	32.43	34.20	33.60	28.10	32.65	31.40	
Ave. Word Length	3.64	3.84	3.84	3.99	3.62	3.98	
Sentences	1,416	57	113	44	143	269	
Sent. length	16.73	20.53	13.30	25.68	17.60	14.93	
sd. Sent. Length	17.64	11.28	6.16	17.08	12.58	9.63	
Paragraphs	708	6	53	10	33	118	
Para. length	44.61	161.00	36.55	95.20	63.45	40.14	
sd. Para. length	83.00	165.68	41.28	148.54	93.22	61.40	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	3,324	114	136	120	240	655	
2-letter words	12,534	468	772	452	1,127	1,607	
3-letter words	2,393	62	196	74	240	270	
4-letter words	3,853	148	238	108	354	380	
5-letter words	2,714	77	125	74	216	477	
6-letter words	3,262	110	176	135	237	523	
7-letter words	1,596	59	87	77	134	188	
8-letter words	1,674	54	100	91	103	315	
9-letter words	737	32	40	13	30	120	

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

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



	N	1	2	3	4	5	6
Text File	OVERALL	UTLWAT-1.TXT	TLHOWA-1	TXTSETSWA-4	TXTSETSWA-3	TXTSETSWA-2	TXTSETSWA-1
Bytes	5,785,661	47,091	37,395	45,330	44,587	78,755	
Tokens	960,040	8,291	6,522	7,043	6,752	11,307	
Types	49,182	1,429	1,330	1,369	1,859	2,839	
Type/Token Ratio	5.12	17.24	20.39	19.44	27.53	25.11	
Standardised Type/Token	35.80	31.84	33.45	35.10	43.08	43.75	
Avg. Word Length	3.76	3.62	3.59	3.97	3.94	3.95	
Sentences	42,386	417	308	71	126	221	
Sent. length	20.01	18.99	20.26	57.73	34.09	38.72	
sd. Sent. length	23.48	13.42	13.01	54.93	30.98	48.45	
Paragraphs	8,167	60	63	84	75	135	
Para. length	115.59	132.63	101.97	83.38	86.23	83.70	
sd. Para. length	751.96	220.06	185.87	163.83	154.29	159.21	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	162,663	1,283	1,138	1,215	1,249	1,997	
2-letter words	308,993	3,014	2,316	1,982	1,879	3,292	
3-letter words	67,974	524	417	582	539	878	
4-letter words	103,088	1,004	697	797	622	1,042	
5-letter words	73,658	650	461	613	570	927	
6-letter words	88,794	644	519	582	646	1,106	
7-letter words	47,805	359	360	311	362	616	
8-letter words	47,463	381	240	355	379	612	
9-letter words	21,061	173	135	257	105	312	

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

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

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

Botswana corpus v/s South African oral corpus

Rank	Word	Rank	Word
8	KE	1	KE



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

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

If one compares the top 100 items in the Botswana text corpus and the South African oral corpus, it is important to observe that 68% of the items in the South African oral corpus are retained while 32% of the items are thrown outside the top 100 items in the Botswana text corpus. If one considers the ousted items in Table 10, one concludes that 29 of the 32 items are still very high falling in the range 101-861 while 2 of these ousted items like *thusa* and *tlaa* falling too far outside the top 100 items. It is

important to note that the Botswana people use more spoken variants than the written ones. For example, they pronounce the future morpheme *tla* as *tlaa* (will) and *e tla* as *tlaa* (come) which is in line with the revised Setswana spelling rules while the South African people use the future morpheme *tla* more frequently in both the spoken and the written corpus. Thus the item *tlaa* is more frequently used in the Botswana corpus and very seldom in the South African written corpus. Given the information above, it is important to conclude that the top 100 items in the Botswana text corpus and the South African oral corpus differ substantially.

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

South African oral corpus v/s Botswana corpus

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



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

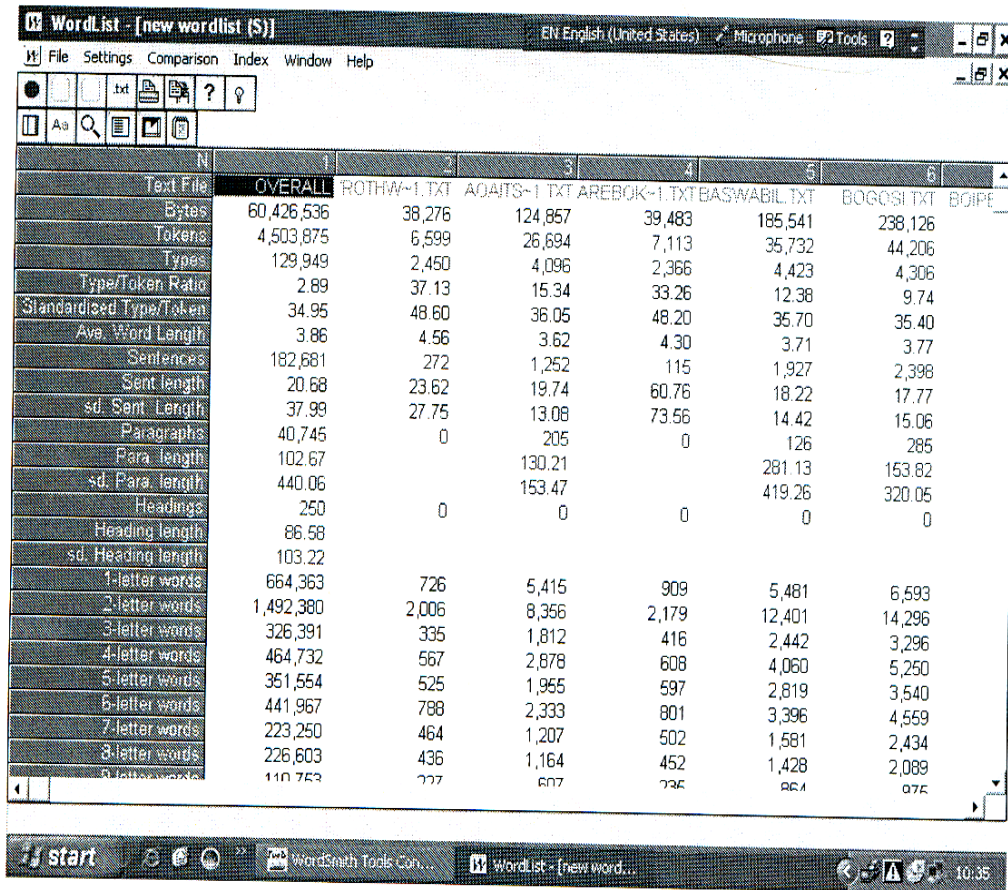
82	MODIMO	278	<u>MODIMO</u>
49	SETSWANA	283	<u>SETSWANA</u>
98	DIANE	333	<u>DIANE</u>
55	ELE	366	<u>ELE</u>
83	NGWAGA	399	<u>NGWAGA</u>
61	BE	405	<u>BE</u>
46	KGWEDI	420	<u>KGWEDI</u>
26	KO	452	<u>KO</u>
53	GONA	489	<u>GONA</u>
99	MALOME	688	<u>MALOME</u>
100	MOTSWANA	861	<u>MOTSWANA</u>
79	DITLHAKA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

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

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

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

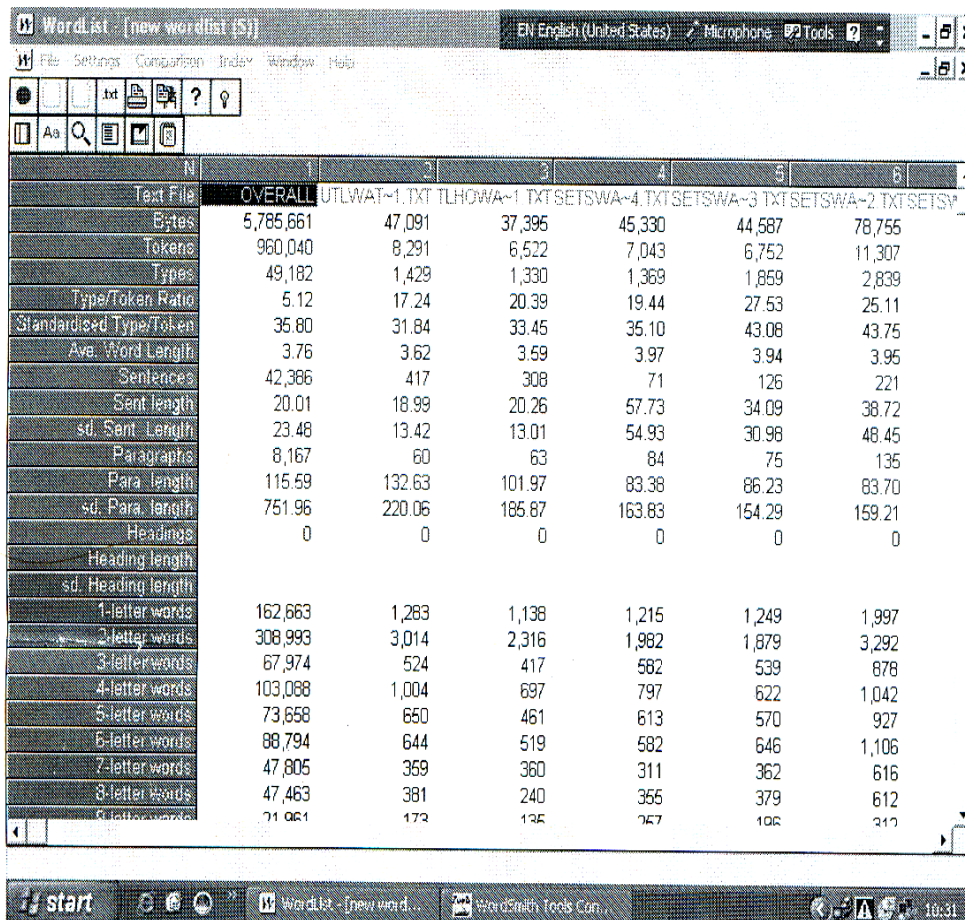
Figure 14 represents the statistical analysis of the South African text corpus in WordSmith Tools comprising of 4,503,875 tokens. Consider the following in this regard:



Text File	N	1	2	3	4	5	6
Bytes	60,426,536	38,276	124,857	39,483	185,541	238,126	
Tokens	4,503,875	6,599	26,694	7,113	35,732	44,206	
Types	129,949	2,450	4,096	2,366	4,423	4,306	
Type/Token Ratio	2.89	37.13	15.34	33.26	12.38	9.74	
Standardised Type/Token	34.95	48.60	36.05	48.20	35.70	35.40	
Ave. Word Length	3.86	4.56	3.62	4.30	3.71	3.77	
Sentences	182,681	272	1,252	115	1,927	2,398	
Sent length	20.68	23.62	19.74	60.76	18.22	17.77	
sd. Sent Length	37.99	27.75	13.08	73.56	14.42	15.06	
Paragraphs	40,745	0	205	0	126	285	
Para length	102.67		130.21		281.13	153.82	
sd. Para length	440.06		153.47		419.26	320.05	
Headings	250	0	0	0	0	0	
Heading length	86.58						
sd. Heading length	103.22						
1-letter words	664,363	726	5,415	909	5,481	6,593	
2-letter words	1,492,380	2,006	8,356	2,179	12,401	14,296	
3-letter words	326,391	335	1,812	416	2,442	3,296	
4-letter words	464,732	567	2,878	608	4,060	5,250	
5-letter words	351,554	525	1,955	597	2,819	3,540	
6-letter words	441,967	788	2,333	801	3,396	4,559	
7-letter words	223,250	464	1,207	502	1,581	2,434	
8-letter words	226,603	436	1,164	452	1,428	2,089	
9-letter words	110,753	227	607	225	864	975	

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

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



Text File	OVERALL	UTLWAT~1.TXT	TLHOWA~1	TXTSETSWA~4	TXTSETSWA~3	TXTSETSWA~2	TXTSETSWA~1
Bytes	5,785,661	47,091	37,395	45,330	44,587	78,755	
Tokens	960,040	8,291	6,522	7,043	6,752	11,307	
Types	49,182	1,429	1,330	1,369	1,859	2,839	
Type/Token Ratio	5.12	17.24	20.39	19.44	27.53	25.11	
Standardised Type/Token	35.80	31.84	33.45	35.10	43.08	43.75	
Ave. Word Length	3.76	3.62	3.59	3.97	3.94	3.95	
Sentences	42,386	417	308	71	126	221	
Sent length	20.01	18.99	20.26	57.73	34.09	38.72	
sd. Sent length	23.48	13.42	13.01	54.93	30.98	48.45	
Paragraphs	8,167	60	63	84	75	135	
Para length	115.59	132.63	101.97	83.38	86.23	83.70	
sd. Para length	751.96	220.06	185.87	163.83	154.29	159.21	
Headings	0	0	0	0	0	0	
Heading length							
sd. Heading length							
1-letter words	162,663	1,263	1,138	1,215	1,249	1,997	
2-letter words	308,993	3,014	2,316	1,982	1,879	3,292	
3-letter words	67,974	524	417	582	539	878	
4-letter words	103,088	1,004	697	797	622	1,042	
5-letter words	73,658	650	461	613	570	927	
6-letter words	88,794	644	519	582	646	1,106	
7-letter words	47,805	359	360	311	362	616	
8-letter words	47,463	381	240	355	379	612	
9-letter words	21,061	173	135	257	106	312	

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

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

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

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

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



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

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

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

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



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

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

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

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

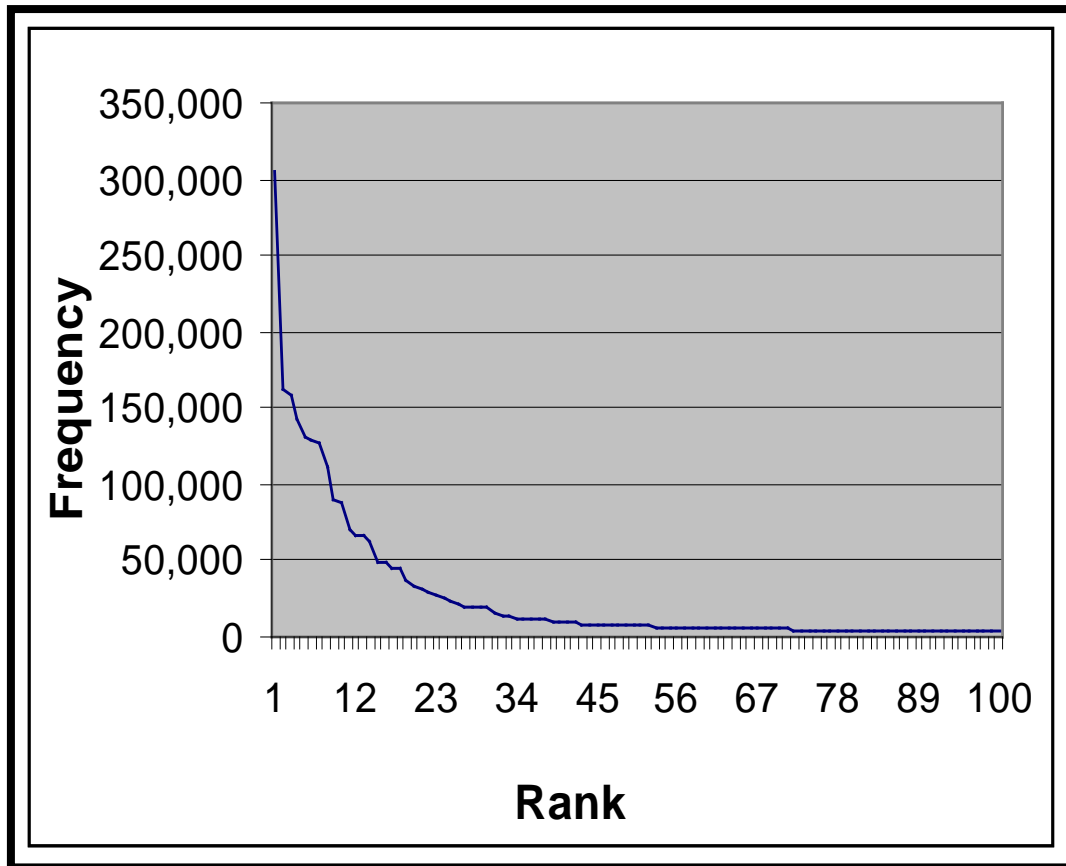
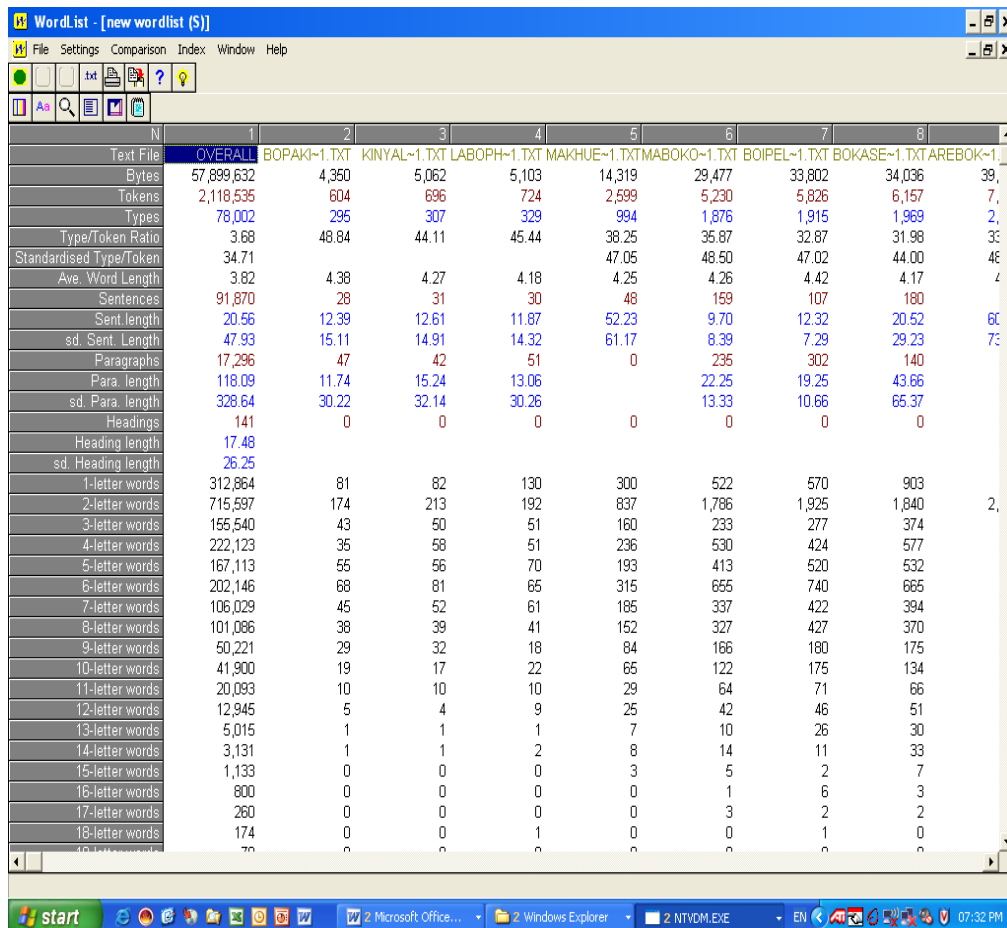


Figure 16: A rapid frequency decline in the top 100 words

3.11 Testing the stability of the Setswana corpus

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

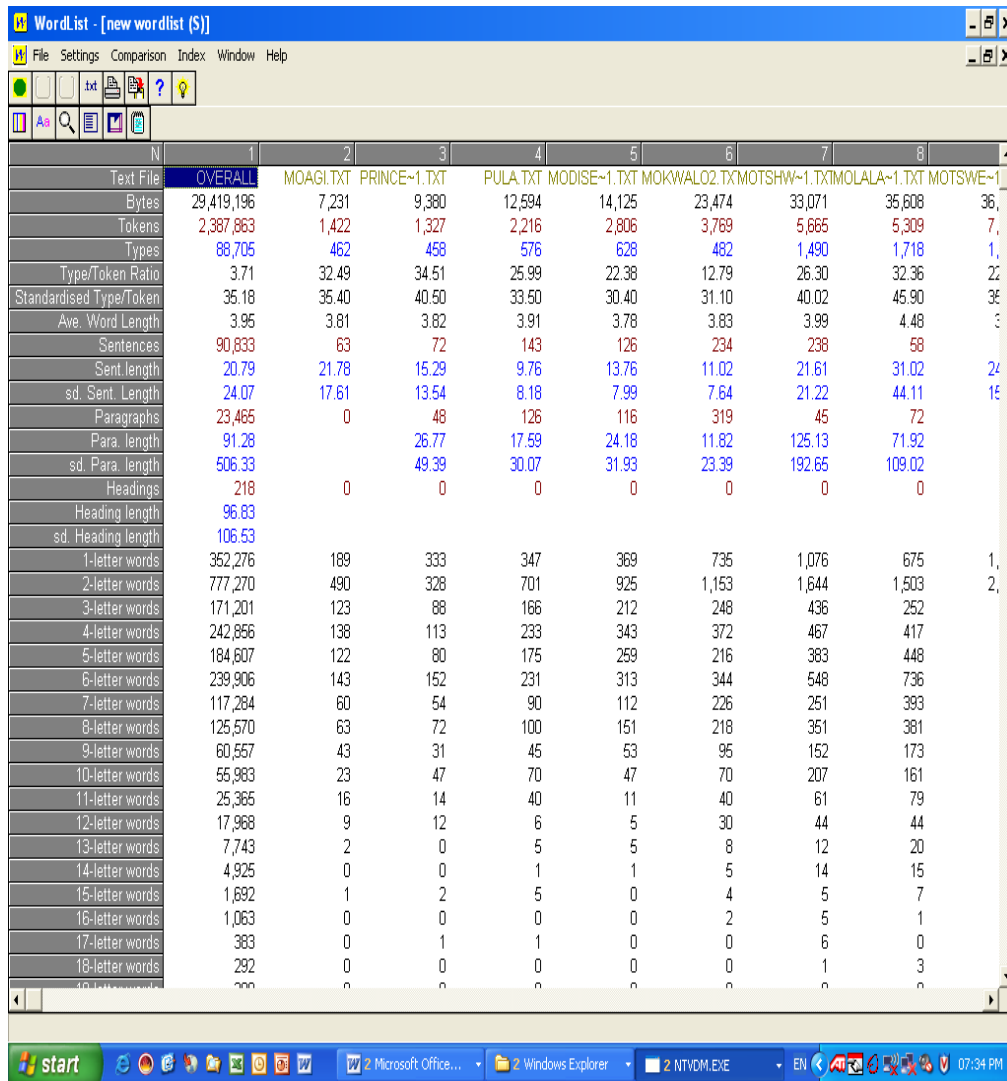
The statistical analysis drawn from the Pretoria Setswana text Corpus (PSetC) will now be used to monitor the growing stages of the Setswana corpus. The first phase (PSetC-Phase1) was doubled to contain 2118, 535 tokens (PSetC-Half1). Consider figure 17 below:



N	1	2	3	4	5	6	7	8	9
Text File	OVERALL	BOPAKI-1.TXT	KINYAL-1.TXT	LABOPH-1.TXT	MAKHUE-1.TXT	MABOKO-1.TXT	BOIPEL-1.TXT	BOKASE-1.TXT	AREBOK-1.TXT
Bytes	57,899,632	4,350	5,062	5,103	14,319	29,477	33,802	34,036	39,111
Tokens	2,118,535	604	696	724	2,599	5,230	5,826	6,157	7,111
Types	78,002	295	307	329	994	1,876	1,915	1,969	2,111
Type/Token Ratio	3.68	48.84	44.11	45.44	38.25	35.87	32.87	31.98	33.05
Standardised Type/Token	34.71				47.05	48.50	47.02	44.00	46.00
Ave. Word Length	3.62	4.38	4.27	4.18	4.25	4.26	4.42	4.17	4.17
Sentences	91,870	28	31	30	48	169	107	180	180
Sent. length	20.56	12.39	12.61	11.87	52.23	9.70	12.32	20.52	60.00
sd. Sent. Length	47.93	15.11	14.91	14.32	61.17	8.39	7.29	29.23	75.00
Paragraphs	17,296	47	42	51	0	235	302	140	140
Para. length	118.09	11.74	15.24	13.06		22.25	19.25	43.66	
sd. Para. length	328.64	30.22	32.14	30.26		13.33	10.66	65.37	
Headings	141	0	0	0	0	0	0	0	
Heading length	17.48								
sd. Heading length	26.25								
1-letter words	312,864	81	82	130	300	522	570	903	
2-letter words	715,597	174	213	192	837	1,786	1,925	1,840	2,111
3-letter words	155,540	43	50	51	160	233	277	374	
4-letter words	222,123	35	58	51	236	530	424	577	
5-letter words	167,113	55	56	70	193	413	520	532	
6-letter words	202,146	68	81	65	315	655	740	665	
7-letter words	106,029	45	52	61	185	337	422	394	
8-letter words	101,086	38	39	41	152	327	427	370	
9-letter words	50,221	29	32	18	84	166	180	175	
10-letter words	41,900	19	17	22	65	122	175	134	
11-letter words	20,093	10	10	10	29	64	71	66	
12-letter words	12,945	5	4	9	25	42	46	51	
13-letter words	5,015	1	1	1	7	10	26	30	
14-letter words	3,131	1	1	2	8	14	11	33	
15-letter words	1,133	0	0	0	3	5	2	7	
16-letter words	800	0	0	0	0	1	6	3	
17-letter words	260	0	0	0	0	3	2	2	
18-letter words	174	0	0	1	0	0	1	0	

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

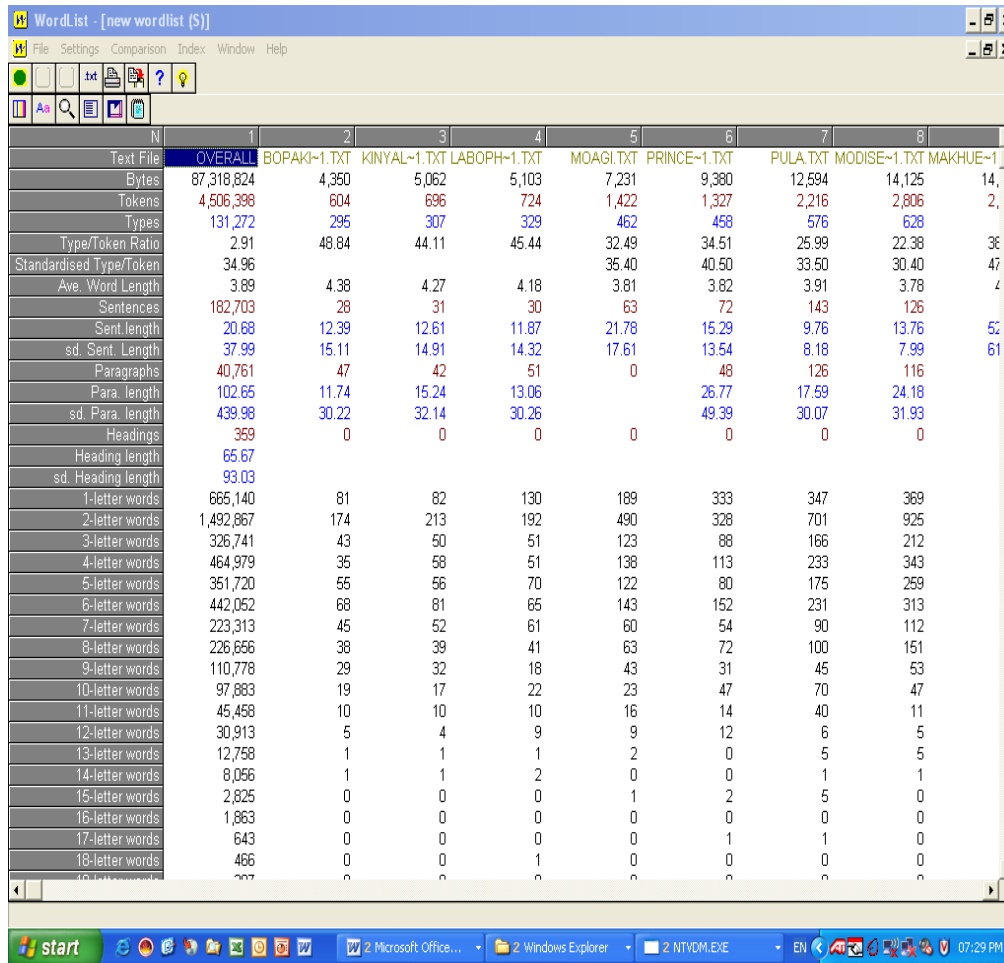
A different corpus was subsequently compiled from a half corpus of 2,387,863 (PSetC-Half 2) tokens. Consider the statistical analysis in figure 18 below:



	N	1	2	3	4	5	6	7	8
Text File	OVERALL	MOAGI.TXT	PRINCE~1.TXT	PULA.TXT	MODISE~1.TXT	MOKWALO2.TXT	MOTSHW~1.TXT	XIMOLALA~1.TXT	MOTSWE~1.TXT
Bytes	29,419,196	7,231	9,380	12,594	14,125	23,474	33,071	35,608	36,100
Tokens	2,387,863	1,422	1,327	2,216	2,806	3,769	5,665	5,309	7,100
Types	88,705	462	458	576	628	482	1,490	1,718	1,718
Type/Token Ratio	3.71	32.49	34.51	25.99	22.38	12.79	26.30	32.36	22.36
Standardised Type/Token	35.18	35.40	40.50	33.50	30.40	31.10	40.02	45.90	38.00
Ave. Word Length	3.95	3.81	3.82	3.91	3.78	3.83	3.99	4.48	3.95
Sentences	90,833	63	72	143	126	234	238	58	58
Sent. length	20.79	21.78	15.29	9.76	13.76	11.02	21.61	31.02	24.00
sd. Sent. Length	24.07	17.61	13.54	8.18	7.99	7.64	21.22	44.11	15.00
Paragraphs	23,465	0	48	126	116	319	45	72	72
Para. length	91.28		26.77	17.59	24.18	11.82	125.13	71.92	71.92
sd. Para. length	506.33		49.39	30.07	31.93	23.39	192.65	109.02	109.02
Headings	218	0	0	0	0	0	0	0	0
Heading length	96.83								
sd. Heading length	106.53								
1-letter words	352,276	189	333	347	369	735	1,076	675	1,076
2-letter words	777,270	490	328	701	925	1,153	1,644	1,503	2,153
3-letter words	171,201	123	88	166	212	248	436	252	252
4-letter words	242,856	138	113	233	343	372	467	417	417
5-letter words	184,607	122	80	175	259	216	383	448	448
6-letter words	239,906	143	152	231	313	344	548	736	736
7-letter words	117,284	60	54	90	112	226	251	393	393
8-letter words	125,570	63	72	100	151	218	351	381	381
9-letter words	60,557	43	31	45	53	95	152	173	173
10-letter words	55,983	23	47	70	47	70	207	161	161
11-letter words	25,365	16	14	40	11	40	61	79	79
12-letter words	17,968	9	12	6	5	30	44	44	44
13-letter words	7,743	2	0	5	5	8	12	20	20
14-letter words	4,925	0	0	1	1	5	14	15	15
15-letter words	1,692	1	2	5	0	4	5	7	7
16-letter words	1,063	0	0	0	0	2	5	1	1
17-letter words	383	0	1	1	0	0	6	0	0
18-letter words	292	0	0	0	0	0	1	3	3
19-letter words	200	0	0	0	0	0	0	0	0

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

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



N	1	2	3	4	5	6	7	8	
Text File	OVERALL	BOPAKI-1.TXT	KINYAL-1.TXT	LABOPH-1.TXT	MOAGI.TXT	PRINCE-1.TXT	PULA.TXT	MODISE-1.TXT	MAKHUE-1
Bytes	87,318,824	4,350	5,062	5,103	7,231	9,380	12,594	14,125	14,125
Tokens	4,506,398	604	696	724	1,422	1,327	2,216	2,806	2,806
Types	131,272	295	307	329	462	458	576	628	628
Type/Token Ratio	2.91	48.84	44.11	45.44	32.49	34.51	25.99	22.38	22.38
Standardised Type/Token	34.96				35.40	40.50	33.50	30.40	30.40
Ave. Word Length	3.89	4.38	4.27	4.18	3.81	3.82	3.91	3.78	3.78
Sentences	182,703	28	31	30	63	72	143	126	126
Sent. length	20.68	12.39	12.61	11.87	21.78	15.29	9.76	13.76	13.76
sd. Sent. Length	37.99	15.11	14.91	14.32	17.61	13.54	8.18	7.99	7.99
Paragraphs	40,761	47	42	51	0	48	126	116	116
Para. length	102.65	11.74	15.24	13.06		26.77	17.59	24.18	24.18
sd. Para. length	439.98	30.22	32.14	30.26		49.39	30.07	31.93	31.93
Headings	369	0	0	0	0	0	0	0	0
Heading length	65.67								
sd. Heading length	93.03								
1-letter words	665,140	81	82	130	189	333	347	369	369
2-letter words	1,492,867	174	213	192	490	328	701	925	925
3-letter words	326,741	43	50	51	123	88	166	212	212
4-letter words	464,979	35	58	51	138	113	233	343	343
5-letter words	351,720	55	56	70	122	80	175	259	259
6-letter words	442,052	68	81	65	143	152	231	313	313
7-letter words	223,313	45	52	61	80	54	90	112	112
8-letter words	226,666	38	39	41	63	72	100	151	151
9-letter words	110,778	29	32	18	43	31	45	53	53
10-letter words	97,883	19	17	22	23	47	70	47	47
11-letter words	45,458	10	10	10	16	14	40	11	11
12-letter words	30,913	5	4	9	9	12	6	5	5
13-letter words	12,758	1	1	1	2	0	5	5	5
14-letter words	8,056	1	1	2	0	0	1	1	1
15-letter words	2,825	0	0	0	1	2	5	0	0
16-letter words	1,863	0	0	0	0	0	0	0	0
17-letter words	643	0	0	0	0	1	1	0	0
18-letter words	466	0	0	1	0	0	0	0	0
19-letter words	207	0	0	0	0	0	0	0	0

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

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

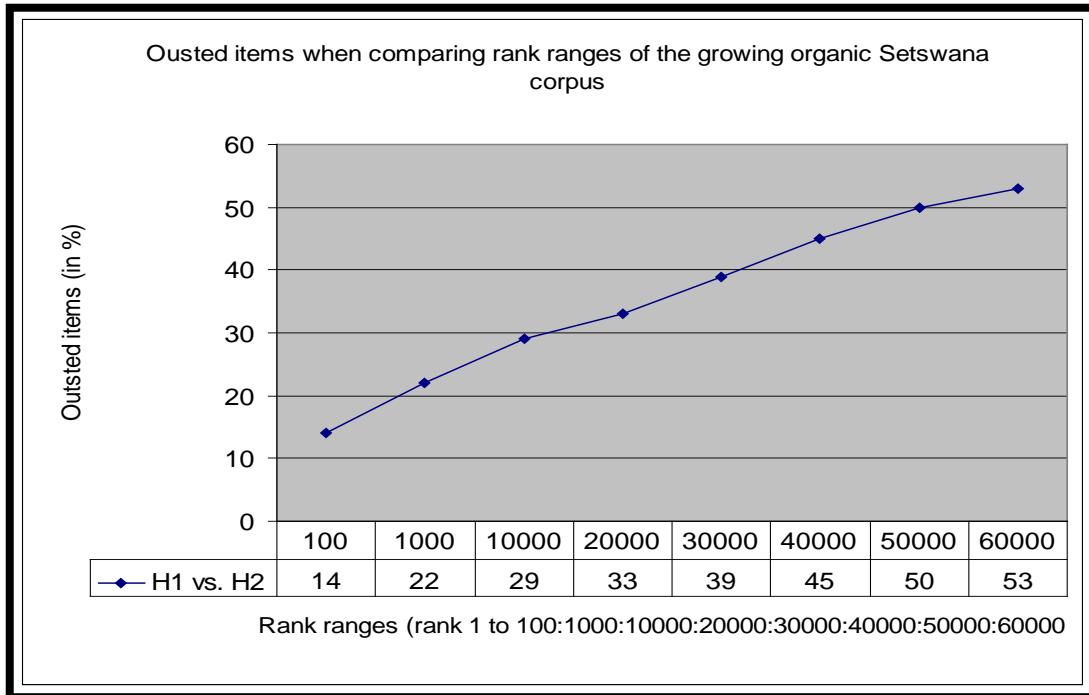


Figure 20: Items when comparing rank ranges of PSetC-Half1 with PSetC-Half2

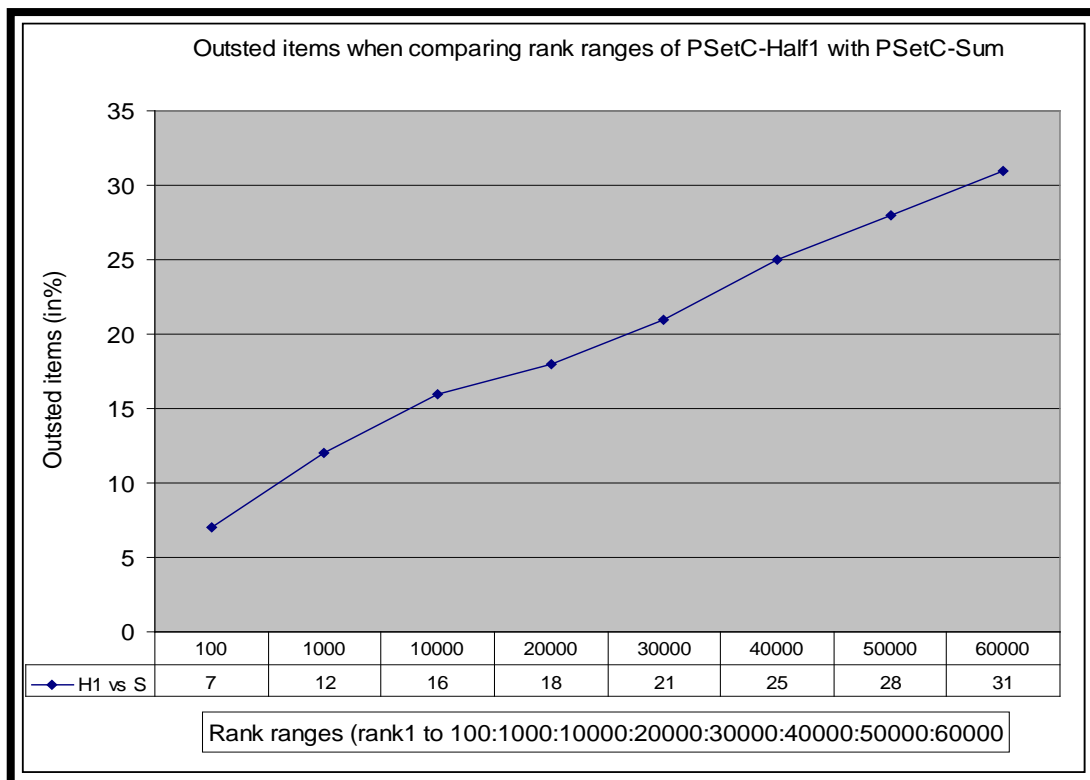


Figure 21: Items when comparing rank ranges of PSetC-Half1 with PSetC-Sum

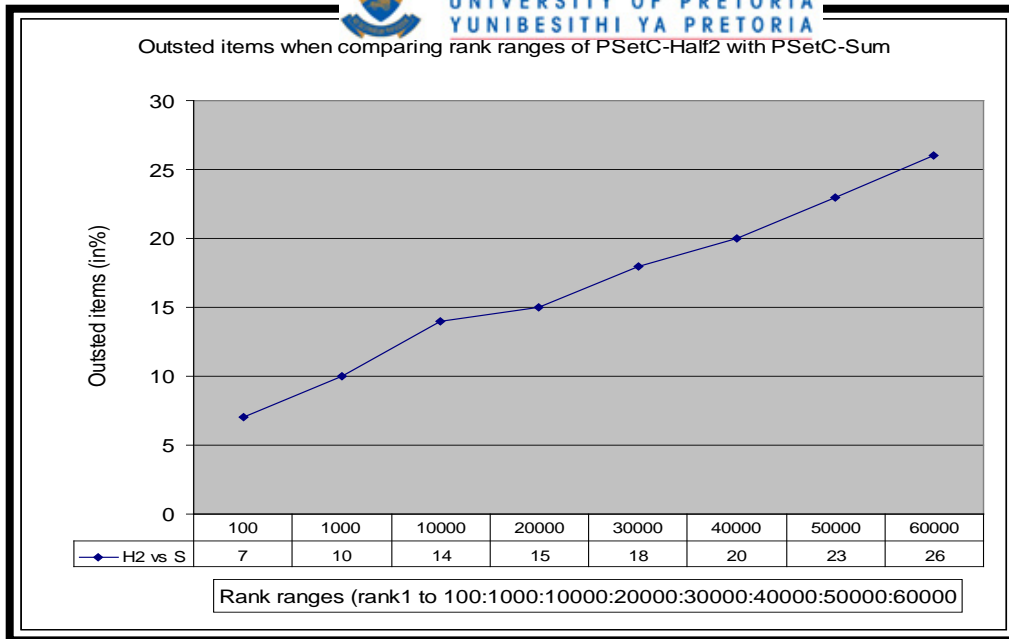


Figure 22: Ousted items when comparing rank ranges of PSetC-Half2 with PSetC-Sum

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

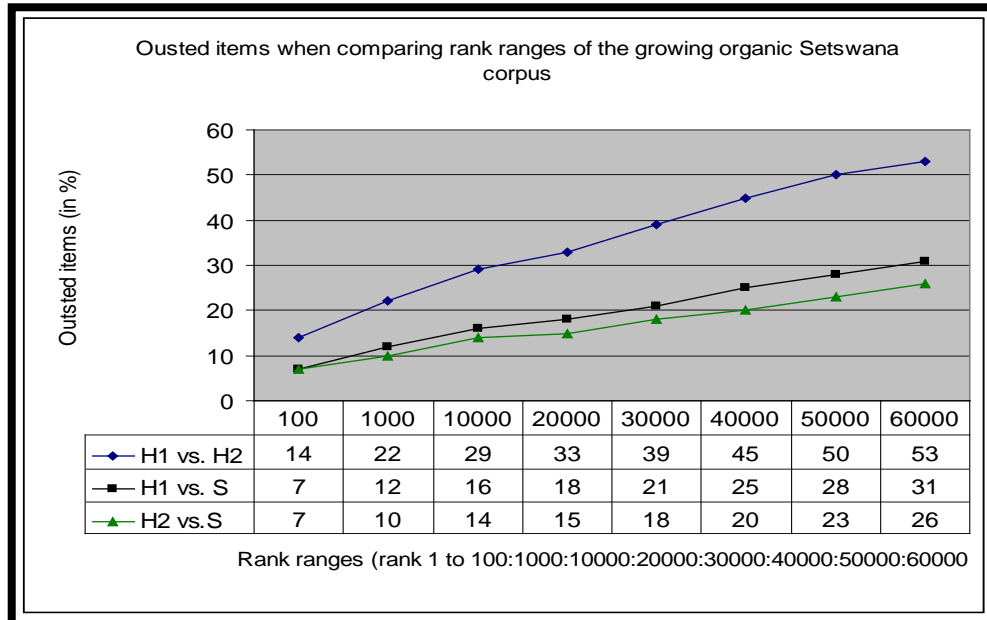


Figure 23: Ousted items when comparing rank ranges of the growing organic Setswana corpus

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

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

Table 14: Summary of SetH1 versus SetH2

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

3.12 Conclusion

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

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

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