Chapter 7

Type/token measures of corpus chunks

7.1 Type/token measures

This chapter measures the degree to which with every additional 10,000 tokens the number of word types grows. Type/token ratio measures lexical richness and determines lexical closure in a text or corpus. If the number of types grows with the addition of every 10,000 tokens it will show that a text has not reached lexical closure. If on the other hand the types do not grow, it will signal lexical closure. What is investigated is the degree to which types grow at comparable points since we seek to determine the lexical richness of different text types at comparable points. The question has been suggested differently by Kjellmer:

> Another method of measuring the density of a text type could be to try and answer the question: How many words (types) has the writer introduced into his text after 100 running words (tokens), how many after 200, etc? The more types he has introduced, the more varied his style is likely to be (Kjellmer, 1994: 117).

The aim is therefore to investigate how types grow at comparable token points in different text types. The purpose of the experiments is to establish whether text types vary in lexical density. The diversity of lexical richness found in genres and domains is crucial for the application of lexicography since a dictionary that aims to capture the language variability will be enriched by a corpus comprising texts from diverse sources.

The results of type measures experiments at comparable points are then plotted in a graph to graphically reveal the text types with both high and low text type growth. The experiments are significant in that they measure word types in different text types at similar numerical intervals making it possible to make useful comparisons between text types.

Statistical studies of vocabulary usually report the ratio between types and tokens for a given sample of text (Baayen, 2001). However such statistics are rarely informative since as more word repetitions occur, the type-token ratio falls regardless of the text studied. The TTR is bound to decline towards zero as tokens increase.

Because of such a phenomenon Youmans argues that:

> …this ratio cannot distinguish any text (or any author) from any other. It is not type-token ratios that are significant, *but only the rate at which they decline.* …Type-token ratios are meaningless, then, unless we also specify the number of tokens used in computing them… But this makes it pointless to compute a ratio at all, since this ratio provides no more information than the raw data do… That is, we can compare the number of types directly rather than the type-token ratios and the ratio between these two pairs of statistics is necessarily the same … it is preferable to plot the number of types in a passage directly against the number of tokens, rather than type-token ratios (Youmans, 1990: 588, italics mine).

Following Youmans, in this study the number of types is plotted directly against the number of tokens of various text types at comparable points. The corpus text types are divided into fifty 10,000 tokens chunks. That is, although other text types have many tokens that could exceed fifty 10,000 word chunks, for these experiments we use only 50 chunks (500,000 tokens). Some text types such as Science and Business have fewer tokens comparatively. As the smallest text types they each has 140,000 tokens and 100,000 tokens respectively, therefore their tokens fail to reach the 500,000 tokens measurement. Although they are smaller comparatively, they are still large enough to be used for useful comparisons.

Measuring series of text chunks at comparable tokens for word types is however sensitive to the order in which the texts (i.e. 10,000 word corpus chunks) are ordered. This is problematic since every experiment repetition is likely to give different results depending on which one of the 10,000 token-chunks was analysed first.

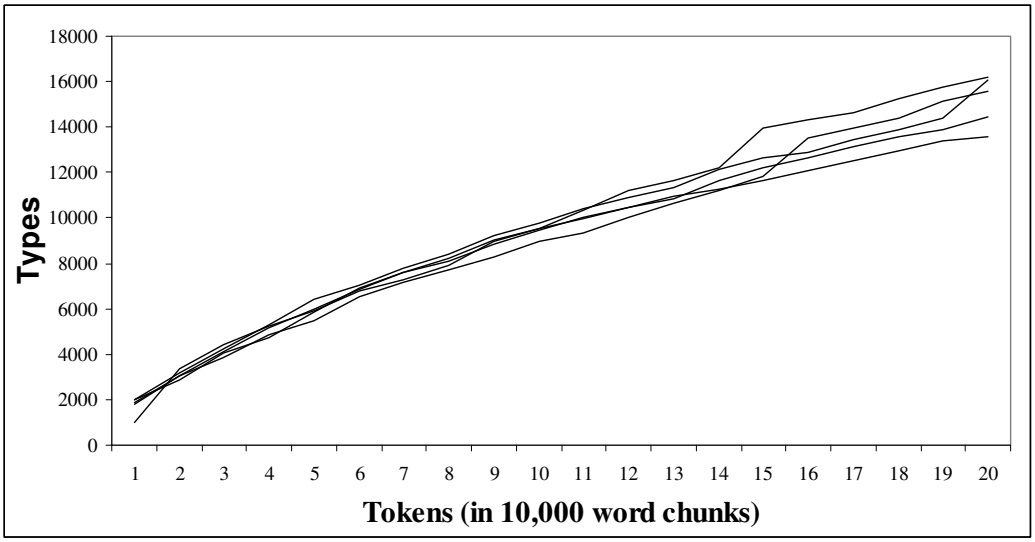
We illustrate this matter below with the five experiment measurements of types from newspaper at 10,000 token intervals up to 200,000 tokens.

Table 52: Newspaper types at 10,000 word tokens intervals

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5
10000	1986	1889	1835	1986	1021
20000	3165	3025	3063	2893	3348
30000	4232	3873	4119	4069	4394
40000	5311	4865	5151	4728	5249
50000	6398	5454	5953	5872	5913
60000	7049	6554	6833	6927	6759
70000	7759	7161	7585	7616	7301
80000	8382	7695	8110	8244	7932
90000	9194	8257	8840	9050	8955
100000	9790	8954	9476	9541	9548
110000	10405	9360	10011	9940	10354
120000	10909	10024	10443	10437	11219
130000	11366	10637	10948	10860	11625
140000	12166	11193	11253	11647	12227
150000	12629	11857	11651	12236	13934
160000	12863	13497	12075	12632	14319
170000	13443	13923	12541	13124	14647
180000	13878	14411	12943	13558	15288
190000	14401	15119	13364	13883	15786
200000	16071	15585	13572	14480	16206

Although Table 52 shows measurements of types from Newspaper text type, whenever an experiment is repeated with a different 10,000 token chunk, this results with a different word type counts. There is a variability of types of the same size from the same text type at comparable token points. For instance, at 200,000 token points there is the following variability of types: 16071, 15585, 13572, 14480, and 16206. This is apparent particularly in Figure 10.

Figure 10: Newspaper types at 10,000 word tokens intervals



The variability of types results from the fact that different 10,000 token chunks are measured at the different token points.

7.1.1 The Mean calculation

To resolve the bias of sequence, the 10,000 token-chunks are randomised for every measurement taken and the experiment iterated five times. The type measurements are taken at every 10,000 token intervals up to 500,000 tokens, repeated five times and an average computed. This is so that we could make comparisons between text types using a single mark or an average that summarises the results i.e. gives an average of types at every 10,000 tokens interval. We therefore compute the measure of central tendency, for which we have chosen the mean. We calculate the mean of the scores using the following formula:

$$\bar{x} = \frac{\sum x}{n}$$

$\bar{x}$ is used for the sample mean; $\sum$ means "the sum of"; x indicates a score and n is used for the number of sample scores. The symbols $\sum x$ means 'add up all the scores'.
The mean is therefore calculated by adding all the scores and dividing their total by their sample size.

The Table 52 scores can therefore be rendered as a table of means that summarises the scores as follows in Table 53:

Table 53: A table of means for Newspaper types

Tokens	Mean
10000	1743.4
20000	3098.8
30000	4137.4
40000	5060.8
50000	5918
60000	6824.4
70000	7484.4

Tokens	Mean
80000	8072.6
90000	8859.2
100000	9461.8
110000	10014
120000	10606.4
130000	11087.2
140000	11697.2

Tokens	Mean
150000	12461.4
160000	13077.2
170000	13535.6
180000	14015.6
190000	14510.6
200000	15182.8

7.1.2 Confidence Interval (CI) calculation

Rather than choosing a single value for the population mean, we can specify a range of values within which we are confident that the value lies (Hinton, 2004: 69). We choose a level of confidence, usually 95% or 99% level of confidence, and then work out the range of values. A level of confidence is the probability that the interval estimate contains the population parameter (Larson and Farber, 2006: 281). If we choose the 95% confidence interval, we are saying that if we worked out the confidence interval for 100 different samples from a population the 95% of those confidence intervals would contain the population mean.

To calculate the confidence interval within which a sample mean lies, we need to know the critical value, standard deviation and the sample size. For our experiments we use 95% confidence interval level as our critical value.

7.1.3 Standard deviation

The standard deviation is a measure of how widely values (raw scores) are dispersed from their mean. We calculate the sample standard deviation using the following formula:

$$s = \sqrt{\frac{\sum (x - \bar{x})2}{n-1}}$$

The lower case *s* represents standard deviation. $\sum(x - \bar{x})^2$ means 'subtract the mean from each raw score to find the deviation score, then square each deviation score and add them all up'. $n - 1$ is what is known as the nonbiased method based on degrees of freedoms (*df*) – the total number of samples minus one. Degrees of freedom concern the scores that contain new information. Pagano (2001: 292) defines degrees of freedom thus: "The degrees of freedom (df) for any statistic is the number of scores that are free to vary in calculating that statistic." There are *N* degrees of freedom associated with the mean since for any set of scores *N* is given. As we have calculated the sample mean from the sample scores we have used up some of the information in

the scores. The number of scores with new information, the degrees of information, is $n - 1$ (Hinton, 2004: 52). We give the example below of a set of scores, their means and the calculation of their standard deviations.

Table 54: Newspaper type scores with mean and standard deviation scores

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5	Mean	SD
10000	1986	1889	1835	1986	1021	1743.4	409.0114
20000	3165	3025	3063	2893	3348	3098.8	169.9741
30000	4232	3873	4119	4069	4394	4137.4	193.4665
40000	5311	4865	5151	4728	5249	5060.8	252.6108
50000	6398	5454	5953	5872	5913	5918	335.0604
60000	7049	6554	6833	6927	6759	6824.4	186.0371
70000	7759	7161	7585	7616	7301	7484.4	245.493
80000	8382	7695	8110	8244	7932	8072.6	268.7263
90000	9194	8257	8840	9050	8955	8859.2	360.7932
100000	9790	8954	9476	9541	9548	9461.8	308.0101
110000	10405	9360	10011	9940	10354	10014	418.8323
120000	10909	10024	10443	10437	11219	10606.4	464.0666
130000	11366	10637	10948	10860	11625	11087.2	400.1983
140000	12166	11193	11253	11647	12227	11697.2	488.4959
150000	12629	11857	11651	12236	13934	12461.4	904.087
160000	12863	13497	12075	12632	14319	13077.2	861.2184
170000	13443	13923	12541	13124	14647	13535.6	798.8284
180000	13878	14411	12943	13558	15288	14015.6	887.9957
190000	14401	15119	13364	13883	15786	14510.6	964.0256
200000	16071	15585	13572	14480	16206	15182.8	1127.631

Having made calculations of standard deviation and determined to use 95% confidence interval, our sample size is at each 10,000 word-token interval up to 500,000 and we can calculate the confidence interval (CI) for the mean.

What we calculate are the upper and lower limits for the 95% confidence interval. We achieve these by calculating the area under the standard normal curve that equals 95%. The value for this area is ± 1.96. This implies that 95% of the area under the standard normal curve falls within 1.96 standard deviations of the mean. The confidence interval is therefore:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

We then calculate the left and right endpoints (or the upper and lower limits for the

confidence interval) and form the confidence interval this way (Larson and Farber, 2006: 297):

Left endpoint: $\bar{x}$ - E

Right endpoint: $\bar{x} + E$

Interval: $\bar{x} - E < \mu < \bar{x} + E$

$\bar{x}$ is the sample mean, E is the margin of error and μ is population mean. Below are the results of the calculation of the confidence interval.

Table 55: Newspaper type scores with mean, critical value, standard deviation and confidence interval scores

Tokens	Exp1	Exp2	Exp3	Exp4	Exp5	Mean	CV	SD	CI	LOW	UPPER
10000	1986	1889	1835	1986	1021	1743.4	0.05	409.0	8.0	1735.4	1751.4
20000	3165	3025	3063	2893	3348	3098.8	0.05	170.0	2.4	3096.4	3101.2
30000	4232	3873	4119	4069	4394	4137.4	0.05	193.5	2.2	4135.2	4139.6
40000	5311	4865	5151	4728	5249	5060.8	0.05	252.6	2.5	5058.3	5063.3
50000	6398	5454	5953	5872	5913	5918	0.05	335.1	2.9	5915.1	5920.9
60000	7049	6554	6833	6927	6759	6824.4	0.05	186.0	1.5	6822.9	6825.9
70000	7759	7161	7585	7616	7301	7484.4	0.05	245.5	1.8	7482.6	7486.2
80000	8382	7695	8110	8244	7932	8072.6	0.05	268.7	1.9	8070.7	8074.5
90000	9194	8257	8840	9050	8955	8859.2	0.05	360.8	2.4	8856.8	8861.6
100000	9790	8954	9476	9541	9548	9461.8	0.05	308.0	1.9	9459.9	9463.7
110000	10405	9360	10011	9940	10354	10014	0.05	418.8	2.5	10011.5	10016.5
120000	10909	10024	10443	10437	11219	10606.4	0.05	464.1	2.6	10603.8	10609.0
130000	11366	10637	10948	10860	11625	11087.2	0.05	400.2	2.2	11085.0	11089.4
140000	12166	11193	11253	11647	12227	11697.2	0.05	488.5	2.6	11694.6	11699.8
150000	12629	11857	11651	12236	13934	12461.4	0.05	904.1	4.6	12456.8	12466.0
160000	12863	13497	12075	12632	14319	13077.2	0.05	861.2	4.2	13073.0	13081.4
170000	13443	13923	12541	13124	14647	13535.6	0.05	798.8	3.8	13531.8	13539.4
180000	13878	14411	12943	13558	15288	14015.6	0.05	888.0	4.1	14011.5	14019.7
190000	14401	15119	13364	13883	15786	14510.6	0.05	964.0	4.3	14506.3	14514.9
200000	16071	15585	13572	14480	16206	15182.8	0.05	1127.6	4.9	15177.9	15187.7

Table 55 shows the 10,000 token interval iterations of types-counts. They are followed by the mean calculations of the five iterations. CV stands for the critical value which is at 5% or 0.05. SD stands for the standard deviation which is followed by the confidence interval (CI) calculation results and the upper and lower confidence interval limits.

With a 95% confidence interval we are saying that if we worked out the confidence interval for 100 different samples from the newspaper section of the Setswana corpus, the 95% of those confidence intervals would contain the population mean. For instance at 200,000 token-population that interval is between 15177.9 and 15187.7. The confidence interval calculations are preferable since they show the confidence intervals which contain the population mean.

The rest of the experiments in tables, henceforth give scores as means calculated from five randomised iterations.

7.2 Text divisions for experiments

For our experiments we have divided the Setswana corpus into the following major text types from the written section of the corpus and that of the spoken subcorpus.

First we discuss the written part of the corpus:

Table 56: Written subcorpus text types

1. Poetry
2. Grammar
3. Chat-site
4. Plays
5. Prose
6. Science
7. Politics
8. Business
9. Religious
10. Newspaper

Miscellaneous text has been left out from experiments since it comprises text from different sources and it is not expected to offer useful information for text type comparison.

The spoken subcorpus has been divided into two major parts:

1. Hansard

2. Call-in, interview and open-radio programming treated as a single unit.

Religious and sport text have been left out as too small for meaningful comparisons.

Further, 50 samples of 10,000 token-chunks were sampled from different text types and combined into what could be termed a single created "text type". We achieved this by randomly dividing the 12 text types into three groups with each having four different text types and sampled text from each text type randomly. We labelled these groups using the initial three letters of each text type in the group, thus: POEGRACHAPLA (Poetry, Grammar, Chat-site and Plays), PRONEWHANCAL (Prose, Newspaper Hansard & Call-in) and SCIPOLBUSREL (Science, Politics, Business and Religious). These divisions are given in Table 57.

The aim of the experiment is to determine the results of comparing subcorpora containing unrelated text with equal-sized subcorpora containing text from a single genre. We measure TTR at comparable points for both texts. The claim is not that {Science, politics, business & religious}, {prose, newspaper, Hansard & call-in} and {poetry, grammar, chat-site & plays} groupings are related in any linguistic way – rather the claim is to the contrary – that they are unrelated and each contribute some distinct types. Combining text from a variety of sources therefore (as one might indeed do in corpus compilation) we hope should give a higher TTR at comparable points compared to that of distinct subcorpus measures.

Table 57: Three divisions of text types

A (POEGRACHAPLA)	
Poetry	13
Grammar	13
Chat-site	12
Plays	12

B (PRONEWHANCAL)	
Prose	13
Newspaper	13
Hansard	12
Call-in	12

C (SCIPOLBUSREL)	
Science	13
Politics	13
Business	12
Religious	12

POEGRACHAPLA, PRONEWHANCAL and SCIPOLBUSREL have 500,000 tokens each. The 500,000 tokens for each newly grouped "text type" was achieved by sampling 13 x 10,000 from two text types and 12 x 10,000 from the remaining two text types to get a total of 500,000 tokens.

This brings to 15 the total number of texts measured and compared.

Table 58: Fifteen major corpus text types

1. Poetry
2. Grammar
3. Chat-site
4. Plays
5. Prose
6. Newspaper
7. Hansard
8. Call-in
9. Science
10. Politics
11. Business
12. Religious
13. POEGRACHAPLA
14. PRONEWHANCAL
15. SCIPOLBUSREL

These three (POEGRACHAPLA, PRONEWHANCAL and SCIPOLBUSREL) have been compiled to test two things: whether the combination of chunks from a variety of text types results in a higher types count at each 10,000 tokens interval compared to a count from a single text type. Second, using the whole Setswana corpus' most frequent 100 words as a standard against which to compare, we generate a frequency list for each of the 15 text types listed above and compare each of their 100 most frequent words against the most frequent 100 words of the whole corpus. Frequency lists present an attractive way of looking at text for statistical analysis. Kilgarriff (1997a: 233) offers at least three advantages to using frequency lists:

i. When a text or corpus is represented as a frequency list, much information is lost, but the tradeoff is an object that is susceptible to statistical processing.
ii. An advantage of using frequency lists is that there is so much data: two corpora can be compared in respect of thousands of data points (e.g., words).
iii. Word frequency lists are cheap and easy to generate.

The frequency lists will therefore be used to compare text types. The assumption is that lists drawn from texts compiled from a variety of text types will be similar to the

one drawn from the entire Setswana corpus, while the list drawn from a single text type is expected to be less similar.

First, we give the results of types at 10,000 token intervals starting with POEGRACHAPLA, Poetry, Grammar, Chat-site and Plays text types.

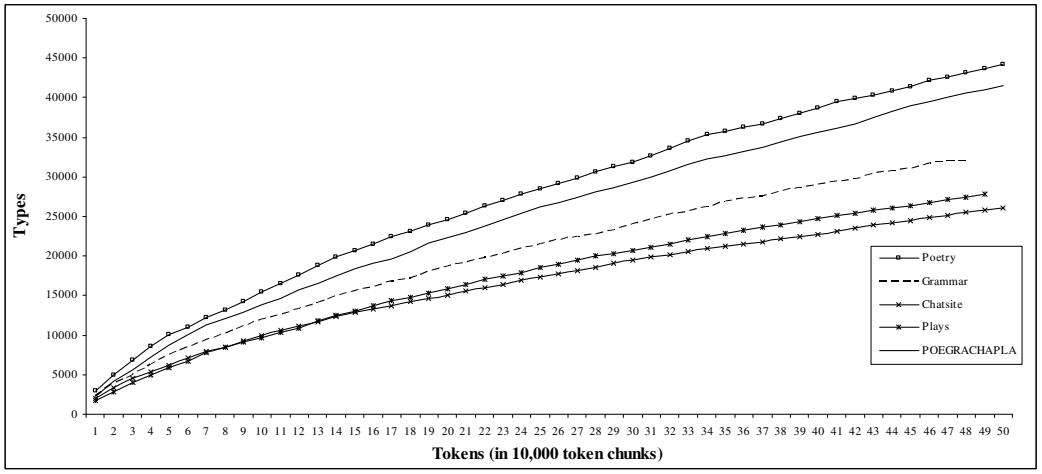
Table 59: Poetry, Grammar, Chat-site, Plays, POEGRACHAPLA text types

Tokens	Poetry	Grammar	Chat-site	Plays	POEGRACHAPLA
10000	2896.6	2274.6	1975.2	1680.2	2223.6
20000	5018	3873.4	3409	2831	4119
30000	6877.4	5022	4528.6	3995	5658
40000	8585.4	6293.8	5323	5015	7210.2
50000	10049.6	7568.2	6121.4	5970.6	8775.6
60000	11057.2	8465	7095.4	6772.2	10127.4
70000	12200.2	9383.6	7882.6	7812.6	11224.6
80000	13222.8	10254.6	8530	8404.2	12120
90000	14241	11157.8	9258.4	9082.6	12884
100000	15494.8	11944.6	9970.6	9683.8	13845.6
110000	16566	12634.2	10587	10321.8	14659.2
120000	17591.4	13292.4	11171	10898.6	15769
130000	18828.4	14085.2	11759.8	11808.8	16595.6
140000	19862	14977.8	12343.8	12475	17528
150000	20677.6	15603.4	12881	13047	18435.8
160000	21509.8	16153.4	13276.4	13762	19109
170000	22399.6	16744	13765	14322.2	19671.6
180000	23134.4	17228	14284.2	14848	20560
190000	23862.8	18022.6	14693.6	15381.8	21645.4
200000	24529.6	18680	15120	15881.2	22354
210000	25341	19275	15634.8	16351.2	22998.8
220000	26308.4	19691.6	16013.2	17004	23764.4
230000	27073.6	20358	16411.2	17502	24568.6
240000	27783.4	21004.4	16959.8	17941.2	25446.8
250000	28517.4	21570.8	17372.2	18487	26267.2
260000	29222.8	22085.2	17808.2	19008	26791.8
270000	29877.6	22474.4	18165.2	19455.6	27434.2
280000	30659	22735.4	18601.8	19974.6	28088.6
290000	31335.6	23263.2	19093.6	20360.4	28621.2
300000	31914.8	24012.4	19426.2	20738.4	29314
310000	32637.2	24644.8	19852.6	21160.4	29939.2
320000	33578.8	25229.8	20224.4	21552.6	30719
330000	34594.2	25724.2	20562.8	22084	31596.2
340000	35307.6	26272.8	20916	22451.4	32216.8
350000	35800.2	26864.4	21214.6	22828.2	32710.2
360000	36300.2	27337	21474.8	23246.6	33254.4
370000	36699.8	27609	21791.6	23605	33800.4
380000	37341.2	28191.8	22111.6	23966.8	34428.8
390000	37977.2	28677	22482.2	24347.8	35037.6
400000	38758	29090.2	22781	24733.4	35626.2

410000	39470.4	29466.2	23121.4	25117.2	36113.4
420000	39922.6	29757.8	23574.6	25460.6	36680
430000	40346.4	30393	23875	25775.4	37544.2
440000	40822.8	30808.8	24180.2	26096.4	38251.2
450000	41352.6	31110.4	24448.2	26410	39023.4
460000	42139.8	31750.4	24828.4	26795.4	39472.8
470000	42656.4	31923.2	25174.2	27094.4	40113.6
480000	43156.4	31989	25478	27416.8	40614
490000	43702.4		25752.8	27819	41029.6
500000	44170.2				41499

The above information is rendered below in graph form. It reveals that poetry has the overall largest number of types.

Figure 11: Prose, Grammar Chat-site, Plays and POEGRACHAPLA types



The graph reveals that from the 10,000 token mark to the 500,000 token point Poetry word types soar above all others. This may offer support to the high lexical density use in poetic language in general. The Poetry text type is followed consistently by POEGRACHAPLA until the end. From 130,000 up to 500,000 tokens Chat-site has the lowest number of types overall. Although Chat-site text has a mixture of Setswana and English words, typos, misspellings, and the general lack of standard spelling, the evidence shows that such language mixture does not translate into high word types.

POEGRACHAPLA has more types than Grammar, Plays and Chat-site texts but lower than Poetry text. The higher level of word types in POEGRACHAPLA suggests that a combination of text from a variety of text types in a corpus may result with

higher levels of types.

Next we measure: Prose, Newspaper, Hansard, Call-in etc (interviews and open radio programs) and PRONEWHANCAL. The results follow in Table 60.

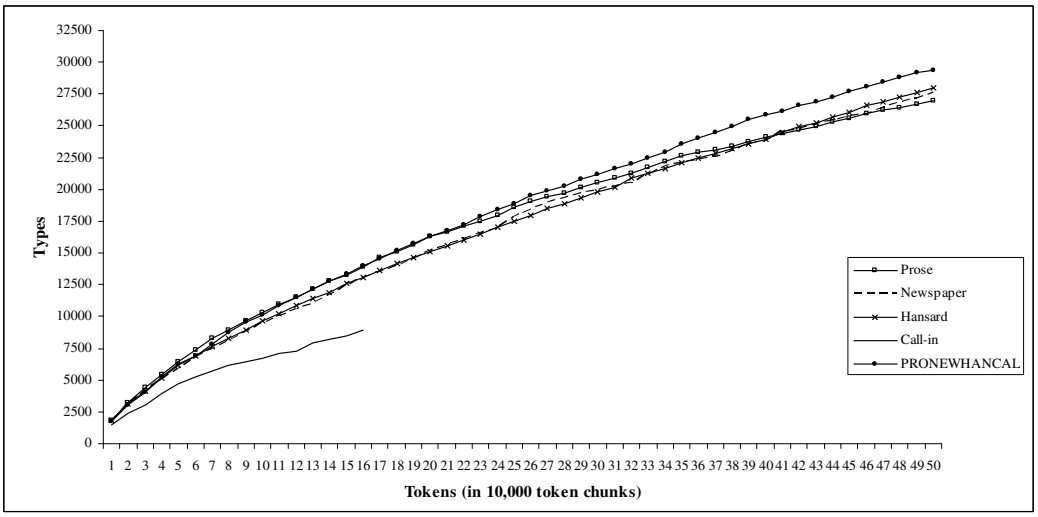
Table 60: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types

Tokens	Prose	Newspaper	Hansard	Call-in, etc	PRONEWHANCAL
10000	1852.6	1743.4	1880	1478.8	1794.6
20000	3216.4	3098.8	3133	2355.8	3076.4
30000	4427.4	4137.4	4145.2	3015.2	4091.8
40000	5430.8	5060.8	5198.8	3951.2	5274.2
50000	6488.8	5918	6178.4	4696	6254
60000	7348.8	6824.4	6929.6	5256.2	6882
70000	8242.8	7484.4	7644.2	5720.2	7781.2
80000	8926.4	8072.6	8328.4	6154.2	8764
90000	9694.8	8859.2	8942.2	6457.8	9543.8
100000	10283.8	9461.8	9644.2	6718.8	10151.8
110000	10939.2	10014	10210.2	7062	10881.8
120000	11477.6	10606.4	10854	7252.6	11501.4
130000	12123.2	11087.2	11404.6	7905.8	12156.4
140000	12783.6	11697.2	11918	8239	12793.4
150000	13242	12461.4	12577.4	8474.2	13378.8
160000	13931.2	13077.2	13108.2	8899	13988
170000	14608	13535.6	13631.6		14570.6
180000	15128.2	14015.6	14152.8		15159.4
190000	15634.6	14510.6	14619.6		15789.2
200000	16254.2	15182.8	15107.8		16270
210000	16686	15672	15587.2		16742.6
220000	17156	16113.2	16012		17257
230000	17465.4	16585.4	16480.4		17880
240000	17970	16981.2	17071		18379
250000	18557.8	17905.6	17462.6		18905.4
260000	19077.4	18372.8	17953.2		19490.8
270000	19408	18945.4	18464.8		19880.6
280000	19724.4	19319.4	18840		20229.4
290000	20181.4	19658	19293.4		20764.6
300000	20510	19960.8	19771.2		21213.4
310000	20919	20227.2	20191.8		21597
320000	21303.8	20540.6	20859.8		22044.2
330000	21731.2	21151	21243		22509.8
340000	22184.8	21780.6	21680		22913
350000	22630	22064.4	22062		23565.2
360000	22930	22340.4	22503.6		23985.8
370000	23138.2	22595.4	22840.2		24517
380000	23392.4	23056.4	23216.8		24925
390000	23795	23539.2	23609.8		25505.4
400000	24138.6	23809.2	23979.4		25884.2

410000	24441.4	24544.6	24518		26171.4
420000	24664.4	24801.8	24993		26568.8
430000	24947.8	25151.6	25266.4		26914
440000	25328	25452.6	25687.6		27252.6
450000	25603.6	25738.8	26073.2		27756.4
460000	25967	25947.8	26599.6		28080.2
470000	26201.8	26393.2	26915.6		28480
480000	26401.4	26853.8	27293.4		28863
490000	26716.2	27115.6	27622.6		29220.4
500000	26941	27607	28005		29367.4

Figure 12 renders the Table 60 results in a graphical form.

Figure 12: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types



From the beginning to the end, Call-in (interview and open radio program) display the lowest number of types compared to Prose, Newspaper and Hansard. This implies that individuals who call radio stations or are interviewed on radio and television, in general, use a limited vocabulary. Between 10,000 and 400,000 tokens Prose has the largest number of word types, after which Hansard word types lead until the 500,000 tokens point. The Hansard types display consistent increase up to the 500,000 tokens point where they are second to the PRONEWHANCAL types and close to the Newspaper types. This may be expected about Hansard text since Hansards document parliamentary debates which are on a variety of topics. The Newspaper types are the most unstable. At certain points they exceed the Hansard types and by the 500,000 tokens point they had exceeded the Prose types. Between 250,000 and 310,000 tokens

they exceed the Hansard types and between 410,000 and 500,000 tokens they exceed the Prose types. From 180,000 tokens PRONEWHANCAL types lead until 500,000 tokens. Since PRONEWHANCAL comprises texts from prose, newspaper, Hansard and call-in text, the high level of types that characterise it, may give support to the view that corpora compiled from a variety of text types have a higher lexical density.

We now turn to Science, Politics, Business, Religious and SCIPOLBUSREL texts. In the entire Setswana corpus, Business, Science and Politics have some of the smallest number of tokens. In terms of our 10,000 chunks they each have 100,000, 140,000 and 200,000 tokens respectively. Religious texts have 480,000 tokens. Below are the results of the calculation of word types for the five text types at comparable points.

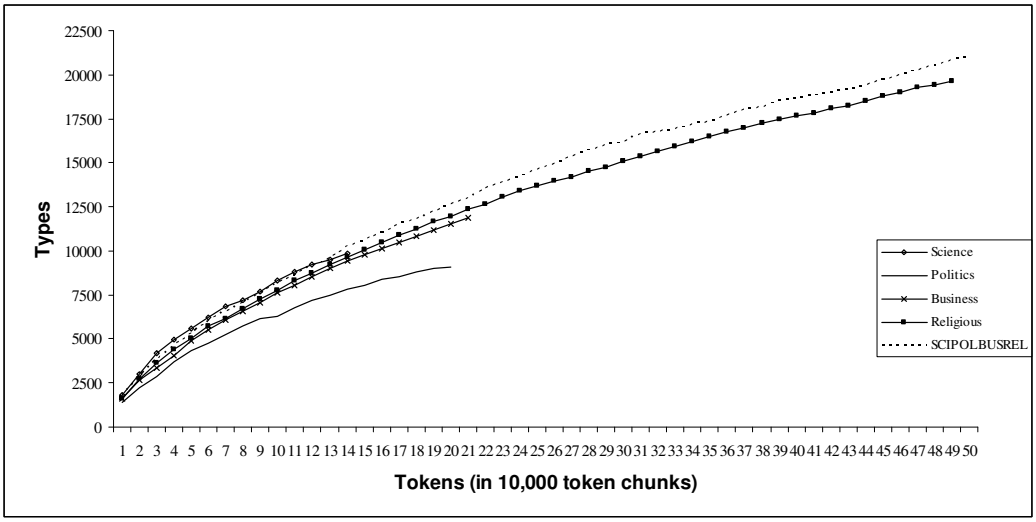
Table 61: Science, Politics, Business, Religious and SCIPOLBUSREL types

Tokens	Science	Politics	Business	Religious	SCIPOLBUSREL
10000	1806	1431	1629.2	1574.6	1751
20000	2972.8	2253	2647.4	2723	2864.6
30000	4167.8	2895	3364	3655.8	3811.4
40000	4970.4	3672	4067.6	4395.2	4614.6
50000	5615.4	4314	4874	5061.4	5309.6
60000	6218.2	4774.2	5507.8	5721.6	6004.2
70000	6841.8	5236.6	6046.6	6177.6	6582
80000	7219.6	5755.6	6587	6741.2	7055
90000	7714.4	6133.4	7070.6	7260	7607.2
100000	8281.8	6274.8	7619.6	7733.2	8172
110000	8778.6	6745.8		8326.4	8622.8
120000	9239.4	7213.8		8763.2	9184.8
130000	9517.2	7490		9229.4	9600
140000	9883	7814.4		9644.6	10262.4
150000		8038		10052.6	10643
160000		8357.8		10448	11008.2
170000		8534		10894.8	11498.6
180000		8780.6		11270.4	11809.4
190000		8997.2		11640.6	12253.4
200000		9063		11971.6	12674
210000				12378.2	12996.8
220000				12682.4	13583.6
230000				13051.6	13899.6
240000				13400.8	14259.4
250000				13672.2	14608.8
260000				13961.2	14973.2
270000				14197.8	15377.8
280000				14507.2	15740.6
290000				14771.4	15984.8
300000				15094.2	16210.6

310000				15399.2	16600.2
320000				15682.8	16767.2
330000				15918.6	16922
340000				16179.2	17171.4
350000				16487.6	17351.2
360000				16750.2	17665.6
370000				16988.6	18001.4
380000				17269.2	18152
390000				17499.6	18495.4
400000				17693.4	18658
410000				17833.4	18846.4
420000				18087.8	18976.8
430000				18264	19142.8
440000				18524	19446.2
450000				18818.8	19713.2
460000				18989.4	19968.2
470000				19259.6	20262.2
480000				19418.6	20529
490000				19646	20810.4
					21047

We plot the above information in the graph below.

Figure 13: Science, Politics, Business, Religious and SCIPOLBUSREL types



From 10,000 to 120,000 tokens Science text leads with the highest types after which it is overtaken by SCIPOLBUSREL which maintains the highest number of types until the 500,000 tokens mark. Not enough data however is available to track the development of the Science text up to the 500,000 tokens mark since it has only 140,000 tokens. From 150,000 Religious text type has the second largest number of types until at 500,000 tokens. Of all the text types Politics have the smallest number

of types. Since SCIPOLBUSREL leads with types between 130,000 and 500,000 token points, this may provide evidence that corpora compiled from a variety of text types do render higher levels of word types.

Having looked at Science, Politics, Business, Religious and SCIPOLBUSREL types we now look at the newspaper text type and measure its subcomponents.

7.2.1 Newspaper Components type/token

While we have looked at the genre of Newspaper text as a single unit above, we recognise that it has different components. This position is similar to that of Kovarik who argues that newspaper texts constitute a sublanguage – a version of a natural language which does not display all of the creativity of that natural language. "The newspaper sublanguage can be further constrained by subject matter to divide it into smaller, more manageable subsets" (Kovarik, 2000: 116/117).

The more manageable subsets that we have isolated in the Setswana newspapers are: Arts and Culture, Business, Letters, News and Sport. These are analysed in a similar manner as other components above. Similarly we give the components' types against token chunks at 10,000 token intervals and we subsequently plot these on a graph.

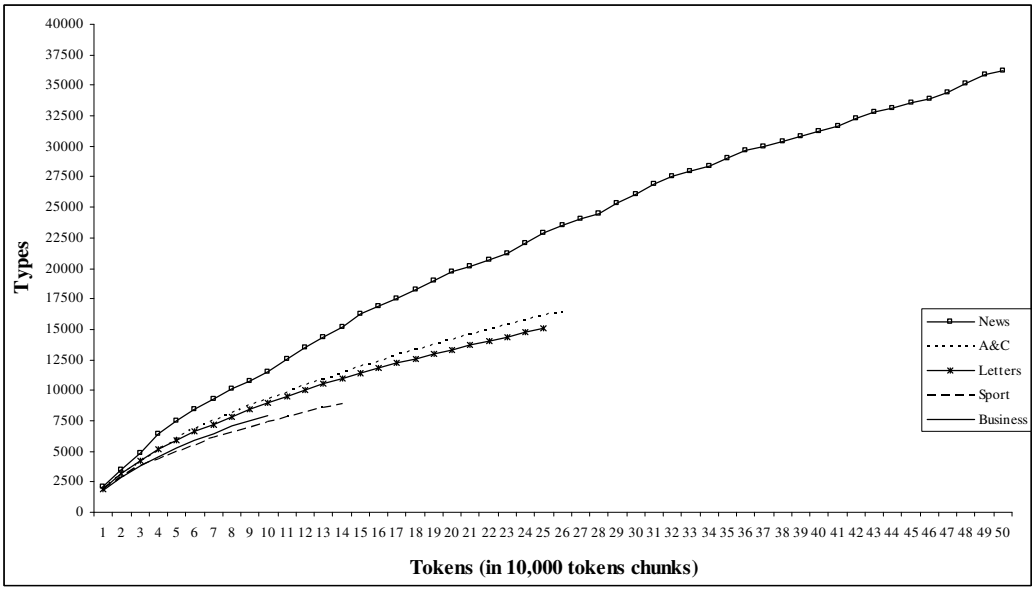
Table 62: Newspaper components types

Tokens	News	A&C	Letters	Sport	Business
10000	2159.2	1927	1901.4	1854.6	1812.2
20000	3530	3140.6	3195.4	2840	2875.8
30000	4886.4	4168.2	4248.4	3795.8	3820.4
40000	6422	5089.2	5151.2	4351.8	4561.4
50000	7516.8	6060.4	5939.6	4955.8	5269.6
60000	8435.8	6808.4	6678.4	5510.6	5922
70000	9239.2	7457.2	7221.2	6090	6459.2
80000	10112.2	8127	7855.6	6562	7028.6
90000	10761.4	8723	8460.4	6992.2	7511.4
100000	11519.4	9276	9021	7418	7937
110000	12554.6	9858.8	9541	7809.4	
120000	13525.2	10414.6	10011.4	8212.8	
130000	14325.2	10892.6	10513.4	8581.6	
140000	15167.4	11399.2	10931	8837.2	
150000	16272.6	11901.8	11350.2		
160000	16921.2	12348.8	11799.4		

170000	17569.8	12899.6	12213.2		
180000	18269.4	13333.2	12573.6		
190000	18944.6	13766.4	12984.8		
200000	19752.2	14134.6	13342.2		
210000	20121	14555.8	13682.2		
220000	20700.8	14989	14057		
230000	21171.4	15389.6	14390.4		
240000	22032.4	15760.6	14762.4		
250000	22940.6	16105	15111		
260000	23537	16392			
270000	24097				
280000	24456.8				
290000	25307				
300000	26055.8				
310000	26863.2				
320000	27503.2				
330000	28018				
340000	28352.6				
350000	28987.6				
360000	29641.2				
370000	29943.2				
380000	30432				
390000	30771.4				
400000	31280.8				
410000	31621.4				
420000	32337.2				
430000	32829.8				
440000	33145.4				
450000	33579.4				
460000	33846.6				
470000	34363.4				
480000	35105.8				
490000	35927.8				
500000	36206				

Table 62 data in graphical form is presented below in Figure 14.

Figure 14: Newspaper components types

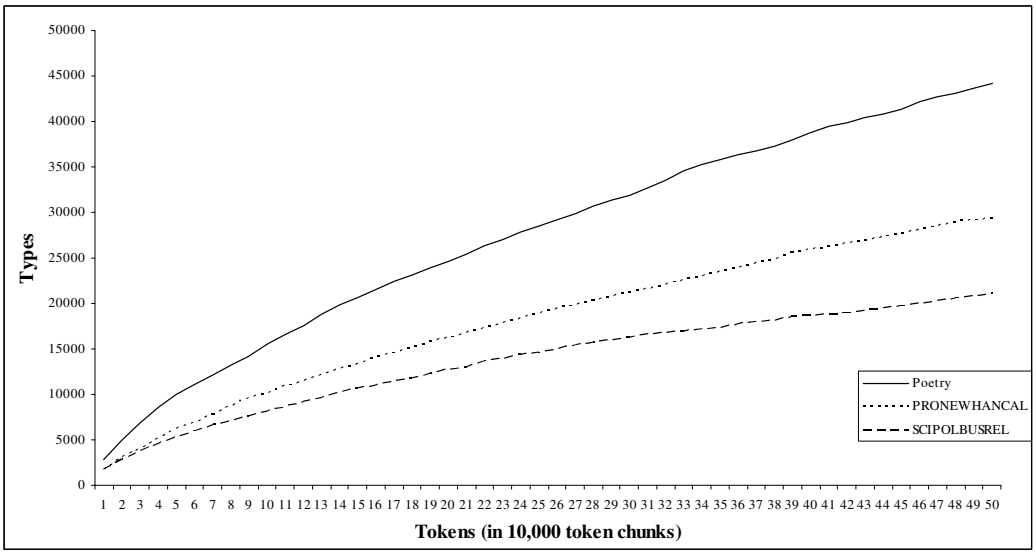


The graph clearly reveals that News types soar above the rest. This is probably because of the different kinds of subjects covered in news compared to Business, Arts and Culture, Letters and Sport. News report on a variety of subjects which we suggest would be responsible for the high number of types compared to the other sections of the newspaper. News word types significantly begin to break away at 20,000 tokens with 3,530 types. Arts and Culture followed by Letters follow News in the number of types, although Letters types are never far removed from the Arts and Culture ones. Sport has the lowest types consistently compared to other text types.

7.3 Conclusion of type-token measurements

In the above experiments we have measured types of various text types at 10,000 tokens intervals. We found that Poetry, PRONEWHANCAL and SCIPOLBUSREL have the largest overall types in general. When we compare these three we find that Poetry leads PRONEWHANCAL and SCIPOLBUSREL. This is reflected in Figure 15 below).

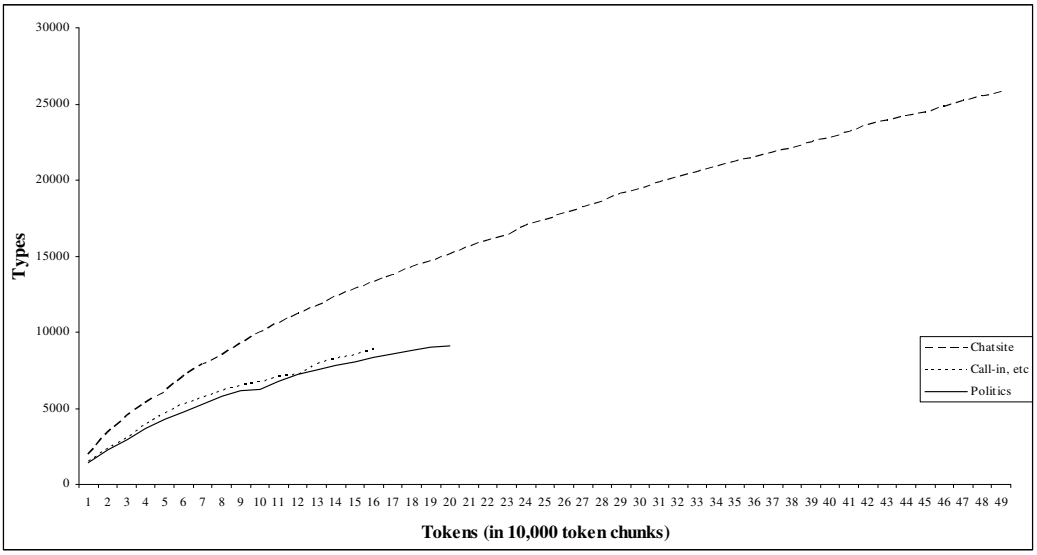
Figure 15: Comparison of the three overall top text types



Overall Poetry text has the largest number of types at most of the 10,000 tokens intervals followed by PRONEWHANCAL and SCIPOLBUSREL respectively. We conclude that poetry uses a wide vocabulary compared to other text types. Given the high number of types in PRONEWHANCAL and SCIPOLBUSREL, and in POEGRACHAPLA, we can safely conclude that combining texts from a variety of text types to compile a corpus leads to a higher number of types.

We have also seen that text types with the lowest types are Chat-site, Call-in and Politics.

Figure 16: Comparison of the three overall lowest text types



Politics have the lowest types overall, followed by Call-in and Chat-site. This suggests that these three use a limited vocabulary when compared with other text types. It should be emphasised, however, that while certain text types contribute the lowest number of types, such text types are not less important or less significant to corpus compilation, since text types with the lowest number of types also do contribute unique words which would enrich a headword list.

Next, we test how frequency lists from different text types and the frequency lists from the three compilations PRONEWHANCAL, SCIPOLBUSREL, and POEGRACHAPLA perform when juxtaposed to the frequency lists generated from the whole corpus.

7.4 A comparison of the top 100 tokens

Below the whole Setswana corpus' most frequent 100 words are used as a standard against which to compare Poetry, Grammar, Chat-site, Plays, Prose, Spoken, Miscellaneous, Science, Politics, Business, Religious, Newspaper, PRONEWHANCAL, SCIPOLBUSREL and POEGRACHAPLA's most frequent 100 words. The purpose of the experiment is to determine the differences between the top 100 words extracted from a mixture of text types (i.e. the whole corpus) and that of

individual corpus text types that form part of the entire corpus. We also wish to determine how the top 100 words of the whole corpus compare to a limited combination of text types as represented in PRONEWHANCAL, SCIPOLBUSREL and POEGRACHAPLA.

There are at least two approaches that could be adopted to extract the 100 frequent tokens from the entire corpus. Raw frequency counts could be ordered from the most frequent to the least frequent. Such an approach's results are in Table 63.

Table 63: Top 100 most frequent tokens in the whole corpus

N	Word	Freq.	Texts
1	a	676,657	2,845
2	go	413,587	2,793
3	e	403,383	2,806
4	le	354,572	2,772
5	o	327,853	2,788
6	ba	311,646	2,699
7	ka	287,741	2,749
8	ke	241,249	2,734
9	ya	225,776	2,752
10	mo	191,304	2,733
11	re	157,637	2,523
12	ga	148,640	2,667
13	fa	141,858	2,630
14	se	131,599	2,540
15	gore	124,504	2,639
16	di	122,905	2,610
17	ne	96,518	2,279
18	wa	94,050	2,613
19	tsa	91,423	2,571
20	sa	80,426	2,569
21	i	71,499	1,385
22	tse	68,069	2,451
23	kwa	67,921	2,432
24	bo	61,587	2,478
25	mme	59,585	2,401
26	tla	54,537	2,191
27	la	48,330	2,383
28	nna	42,931	2,280
29	yo	36,420	2,002
30	fela	36,309	2,203
31	gagwe	34,149	1,634
32	na	32,970	2,107
33	bona	32,779	1,777
34	bone	30,405	2,214
35	jwa	27,876	2,005
36	jaaka	27,285	2,063
37	batho	26,891	1,979
38	the	24,563	771
39	lo	24,264	951
40	itse	23,349	1,537
41	ntse	23,319	1,614
42	motho	21,357	1,572
43	teng	20,581	1,798
44	to	20,405	632
45	mongwe	20,121	1,681
46	neng	19,556	1,593
47	dira	19,365	1,816
48	jalo	18,207	1,868
49	ene	18,031	1,506
50	bua	17,075	1,420
51	tswa	16,847	1,772
52	rona	16,286	1,488
53	me	16,236	878
54	thata	15,630	1,731
55	kgotsa	15,522	1,231
56	pele	14,992	1,633
57	and	14,081	693
58	of	13,823	836
59	morago	13,687	1,528
60	posted	13,636	327
61	gago	13,515	835
62	kana	13,508	1,387
63	jaanong	13,278	1,323
64	eng	13,277	1,192
65	tshwanetse	12,812	1,310
66	bana	12,553	1,147
67	nako	12,270	1,465
68	batla	11,632	1,382
69	you	11,573	413
70	gape	11,563	1,577
71	yone	11,540	1,512

72	madi	11,393	1,290
73	nngwe	11,069	1,268
74	setse	10,905	1,221
75	ngwana	10,883	827
76	monna	10,832	774
77	tsaya	10,692	1,352
78	leng	10,656	1,238
79	bangwe	10,585	1,611
80	gone	10,531	1,417
81	bile	10,477	1,333
82	ntlha	10,323	1,093
83	dilo	10,248	1,296
84	jaana	10,176	1,247
85	wena	10,070	813
86	tsena	10,056	1,079

87	on	10,032	530
88	is	9,982	541
89	rile	9,982	906
90	utlwa	9,929	932
91	be	9,827	773
92	jang	9,816	1,314
93	tiro	9,770	1,170
94	kgosi	9,668	477
95	sengwe	9,623	1,146
96	tota	9,576	1,381
97	jo	9,532	1,299
98	lefatshe	9,438	1,335
99	botswana	9,433	1,502
100	sentle	9,418	1,252

The results of Table 63 are useful and may be used in the compilation of a headword list. The results are listed on the basis of frequency of occurrence in the entire Setswana corpus. *A* is the most frequent token in the corpus occurring 676,657 times and found in 2,845 texts. *Sentle* occupies the 100th word spot with 9,418 occurrences in 1,252 texts. However Leech et al. (2001: 17) contend that "simple word frequency counts can be misleading." This is because,

> If a word has a high frequency count, the user may infer, because the compilers have attempted to build a large, maximally representative corpus, that the word has a similarly high occurrence in the … language as a whole. However this may be a false inference. It is possible that the word has a high frequency not because it is widely used in the language as a whole but because it is 'overused' in a much smaller number of texts, or parts of texts, within the corpus (Leech et al., 2001: 17).

To address this matter they suggest dispersion statistics (Range (Ra) and Dispersion (Disp)) which show whether a word is widely spread because it occurs in many of the text samples or whether it is because of high usage in only a few samples. They argue that,

> Frequent words with high dispersion values may be considered to have high currency in the language as a whole; high frequencies associated with low dispersion values should, in contrast, be treated with caution (Leech et al.,

2001: 18).

We will not explore any further the complexities of Leech et al.'s statistics, but we discussed them since they bare close semblance to Scott's (2004-2006: 109) Simple Consistency Analysis (SCA). SCA calculates words which recur consistently in lots of texts of a given genre and orders them on the basis of their spread. What SCA does is therefore to calculate word spread. SCA results are given on the basis of the number of texts the words occur in. The results are given in the word-list, for instance Table 64, in a column headed "Texts" which shows the calculated number of texts each word occurred in (the maximum number being the total number of text-files used for the word-list).

SCA is dependent on the number of text-files. The words occurring in the largest number of text files are listed at the top, while the ones occurring in fewer texts occur lower in the list. In Table 64 the top 100 words of the Setswana corpus are given on the basis of SCA measurement.

Table 64: Top 100 words: Simple Consistency Analysis results

N	Word	Freq.	%	Texts
1	a	676,657	5.22	2,845
2	e	403,383	3.11	2,806
3	go	413,587	3.19	2,793
4	o	327,853	2.53	2,788
5	le	354,572	2.74	2,772
6	ya	225,776	1.74	2,752
7	ka	287,741	2.22	2,749
8	ke	241,249	1.86	2,734
9	mo	191,304	1.48	2,733
10	ba	311,646	2.40	2,699
11	ga	148,640	1.15	2,667
12	gore	124,504	0.96	2,639
13	fa	141,858	1.09	2,630
14	wa	94,050	0.73	2,613
15	di	122,905	0.95	2,610
16	tsa	91,423	0.71	2,571
17	sa	80,426	0.62	2,569
18	se	131,599	1.02	2,540
19	re	157,637	1.22	2,523
20	bo	61,587	0.48	2,478
21	tse	68,069	0.53	2,451
22	kwa	67,921	0.52	2,432
23	mme	59,585	0.46	2,401
24	la	48,330	0.37	2,383
25	nna	42,931	0.33	2,280
26	ne	96,518	0.74	2,279
27	bone	30,405	0.23	2,214
28	fela	36,309	0.28	2,203
29	tla	54,537	0.42	2,191
30	na	32,970	0.25	2,107
31	jaaka	27,285	0.21	2,063
32	jwa	27,876	0.22	2,005
33	yo	36,420	0.28	2,002
34	batho	26,891	0.21	1,979
35	jalo	18,207	0.14	1,868
36	dira	19,365	0.15	1,816
37	teng	20,581	0.16	1,798
38	bona	32,779	0.25	1,777
39	tswa	16,847	0.13	1,772
40	thata	15,630	0.12	1,731
41	mongwe	20,121	0.16	1,681
42	gagwe	34,149	0.26	1,634
43	pele	14,992	0.12	1,633
44	ntse	23,319	0.18	1,614
45	bangwe	10,585	0.08	1,611
46	neng	19,556	0.15	1,593
47	gape	11,563	0.09	1,577
48	motho	21,357	0.16	1,572
49	itse	23,349	0.18	1,537
50	morago	13,687	0.11	1,528
51	yone	11,540	0.09	1,512

52	ene	18,031	0.14	1,506
53	botswana	9,433	0.07	1,502
54	rona	16,286	0.13	1,488
55	nako	12,270	0.09	1,465
56	bua	17,075	0.13	1,420
57	gone	10,531	0.08	1,417
58	kana	13,508	0.10	1,387
59	i	71,499	0.55	1,385
60	batla	11,632	0.09	1,382
61	tota	9,576	0.07	1,381
62	tsaya	10,692	0.08	1,352
63	one	9,057	0.07	1,343
64	lefatshe	9,438	0.07	1,335
65	bile	10,477	0.08	1,333
66	jaanong	13,278	0.10	1,323
67	jang	9,816	0.08	1,314
68	tshwanetse	12,812	0.10	1,310
69	jo	9,532	0.07	1,299
70	dilo	10,248	0.08	1,296
71	madi	11,393	0.09	1,290
72	nngwe	11,069	0.09	1,268
73	sentle	9,418	0.07	1,252
74	dingwe	8,177	0.06	1,251
75	jaana	10,176	0.08	1,247
76	sena	7,506	0.06	1,243

77	leng	10,656	0.08	1,238
78	kgotsa	15,522	0.12	1,231
79	setse	10,905	0.08	1,221
80	rre	8,383	0.06	1,212
81	nne	8,023	0.06	1,204
82	eng	13,277	0.10	1,192
83	tiro	9,770	0.08	1,170
84	batswana	5,902	0.05	1,167
85	bana	12,553	0.10	1,147
86	sengwe	9,623	0.07	1,146
87	supa	6,033	0.05	1,145
88	gaborone	3,326	0.03	1,128
89	eo	7,017	0.05	1,125
90	ngwaga	5,216	0.04	1,116
91	dumela	4,380	0.03	1,096
92	ntlha	10,323	0.08	1,093
93	tsone	6,274	0.05	1,093
94	tsotlhe	7,330	0.06	1,086
95	tsena	10,056	0.08	1,079
96	gongwe	8,412	0.06	1,053
97	mokgosi	2,704	0.02	1,046
98	sepe	7,045	0.05	1,040
99	seka	3,713	0.03	1,033
100	raya	9,132	0.07	1,018

Table 64 in the first column shows the rank of a word, followed by the word ranked, which is followed by the word's frequency in the whole corpus. The fourth column is of the word's frequency as a percentage of the corpus, followed by the SCA score which is the number of texts each word appears in.

When we compare the SCA results with those of raw frequencies, we find that all the English words that appear in the top 100 raw frequencies words no longer appear in the 100 SCA results. The English words are, *I* (21), *on* (87), *you* (69), *posted* (60), *is* (88), *be* (91), *the* (39), *of* (58). English words are not spread throughout the corpus but are limited to a few files, that is why they do not appear amongst the 100 SCA results.

We will however use raw frequency measure as a standard against which we measure the most frequent words from different text types to make our study comparable to many studies in the field which use raw frequencies and not dispersion results. Additionally, we use raw frequency counts so that later in this chapter we could compare our lists and results with the BNC lists. In practical dictionary compilation, Leech et al.'s guidance that both raw frequencies and dispersion results should be considered in the selection of headwords.

Our study is similar to that of Sharoff (2006) in which he investigates the possibility to develop a BNC-like corpus for a number of different languages (Chinese, English, German, Romanian, Ukrainian and Russian). He also evaluates the collected corpora using the composition of resulted corpora and their frequency lists for some of the languages (English, German and Russian). He compares the internet compiled corpus with large available balanced English and Russian corpora. For English he used the BNC, for Russian, he used the Russian Reference Corpus (RRC). It is particularly the sections on corpus comparison by Sharoff that interest us.

We compare each of the 15 text types' most frequent 100 types against those of the whole corpus' most frequent 100 tokens. The 15 groups are the following:

A	B	C
Poetry	Science	Prose
Grammar	Politics	Newspaper
Chat-site	Business	Hansard
Plays	Religious	Call-in
POEGRACHAPLA	SCIPOLBUSREL	PRONEWHANCAL

Below we give results in Tables 65, 66, and 67 and follow the results with a discussion. Where a word in the top 100 words of the whole corpus is not in the top 100 words of a text type we indicate such absence by "–" followed by a bracketed number to show the rank it occupies in the list.

Table 65: Poetry, Grammar, Chat-site, Plays and POEGRACHAPLA

Whole corpus	Poetry	Grammar	Chat-site	Plays	POEGRACHAPLA
1. a	1	1	4	1	1
2. e	5	3	31	7	5
3. go	4	2	18	4	2
4. o	6	6	14	3	4
5. le	3	4	20	6	6
6. ya	9	7	58	13	10
7. ka	7	5	41	5	7
8. mo	16	9	60	11	12
9. ke	2	11	19	2	3
10. ba	8	8	26	8	8
11. ga	11	15	51	10	9
12. fa	17	13	100	14	15
13. gore	33	18	43	15	19

14. di	12	10	87	18	14
15. wa	14	19	92	16	16
16. tsa	18	17	- (153)	25	22
17. sa	15	20	- (110)	20	17
18. se	13	12	- (123)	12	13
19. re	10	16	34	9	11
20. tse	32	14	- (275)	29	30
21. bo	19	24	74	21	23
22. kwa	24	23	- (327)	22	29
23. mme	25	22	- (216)	24	33
24. la	20	27	- (240)	34	38
25. nna	23	28	93	23	27
26. ne	27	21	- (192)	17	26
27. bone	37	51	- (202)	45	50
28. tla	21	25	- (162)	19	25
29. fela	26	34	- (189)	30	42
30. jaaka	25	30	- (407)	38	39
31. na	31	29	- (233)	33	41
32. jwa	34	32	- (686)	60	55
33. batho	35	38	- (176)	43	48
34. jalo	- (121)	52	- (499)	65	73
35. yo	38	26	- (221)	27	35
36. teng	52	50	- (277)	49	63
37. tswa	45	58	- (460)	54	70
38. dira	81	35	- (456)	57	74
39. thata	58	36	- (591)	61	59
40. bona	28	33	- (203)	28	46
41. mongwe	86	43	- (365)	62	86
42. pele	62	75	- (751)	79	93
43. gagwe	53	45	- (13,519)	36	51
44. bangwe	72	88	- (428)	- (114)	- (128)
45. gape	- (134)	68	- (342)	- (111)	- (109)
46. ntse	54	57	- (402)	37	57
47. neng	- (261)	78	- (1,201)	44	- (108)
48. botswana	- (50)	- (184)	71	- (1,423)	84
49. yone	82	55	- (455)	- (139)	83
50. motho	30	40	- (201)	35	53
51. morago	76	72	- (992)	82	- (106)
52. itse	36	53	- (205)	31	49
53. rona	46	76	- (182)	41	66
54. nako	- (106)	62	- (512)	- (103)	- (104)
55. ene	64	- (105)	- (366)	48	80
56. gone	- (173)	- (108)	- (259)	85	- (124)
57. bua	56	67	- (251)	42	61
58. kana	63	81	- (184)	50	79
59. batla	67	91	- (253)	51	78
60. lefatshe	43	- (202)	- (1,311)	- (131)	- (183)
61. tota	77	- (183)	- (226)	69	- (178)
62. tsaya	- (110)	93	- (695)	89	- (132)
63. madi	- (107)	- (130)	- (423)	87	- (179)
64. one	- (143)	- (112)	53	- (127)	76
65. tshwanetse	- (440)	84	- (782)	83	- (206)
66. bile	- (155)	- (129)	- (726)	72	- (133)

67. jaanong	- (139)	- (121)	- (351)	52	98
68. jang	- (148)	69	- (485)	74	- (105)
69. dilo	- (104)	82	- (445)	91	- (134)
70. nngwe	- (168)	56	- (1,363)	- (151)	- (190)
71. dingwe	- (183)	47	- (1,383)	- (207)	- (176)
72. rre	- (137)	- (212)	- (901)	- (105)	- (205)
73. sentle	- (171)	86	- (558)	70	- (120)
74. jaana	- (321)	- (128)	- (385)	64	100
75. kgotsa	83	31	- (963)	- (130)	- (157)
76. sena	- (329)	- (199)	- (571)	- (172)	- (250)
77. setse	- (103)	- (142)	- (724)	73	- (141)
78. nne	94	99	- (795)	- (126)	- (172)
79. tiro	- (114)	65	- (1,581)	81	- (195)
80. leng	- (150)	80	- (730)	- (112)	- (177)
81. batswana	- (105)	- (384)	86	- (562)	- (142)
82. supa	- (178)	70	- (3,772)	- (206)	- (295)
83. jo	- (215)	- (101)	- (1,108)	- (101)	- (186)
84. gaborone	- (181)	- (651)	- (230)	- (623)	- (228)
85. sengwe	- (142)	49	- (694)	71	- (151)
86. ngwaga	- (191)	- (198)	- (3,467)	- (358)	- (334)
87. eo	- (239)	- (203)	- (931)	- (166)	- (278)
88. bana	41	64	- (279)	80	82
89. tsone	- (129)	- (107)	- (868)	189	- (146)
90. eng	- (144)	42	- (295)	39	52
91. mokgosi	- (304)	- (2,022)	- (28,185)	- (773)	- (772)
92. tsotlhe	73	- (127)	- (2,762)	- (152)	- (182)
93. ntlha	- (113)	60	- (1,996)	- (106)	- (137)
94. dumela	43	- (330)	- (237)	- (117)	- (192)
95. tlhalosa	- (555)	79	- (6,527)	- (520)	- (361)
96. tsena	68	- (191)	- (556)	59	91
97. gongwe	- (125)	74	- (529)	100	- (184)
98. mangwe	- (302)	- (134)	- (2,469)	- (428)	- (309)
99. gompieno	- (257)	- (287)	- (1,006)	- (113)	- (212)
100.seka	- (429)	- (943)	- (491)	- (398)	- (576)
Total	**62**	**74**	**21**	**72**	**59**

Table 66: Science, Politics, Business, Religious and SCIPOLBUSREL

Whole corpus	Science	Politics	Business	Religious	SCIPOLBUSREL
1. a	5	5	1	1	4
2. e	4	4	2	5	3
3. go	1	1	3	4	1
4. o	11	11	5	6	9
5. le	2	2	4	2	2
6. ya	3	3	6	8	5
7. ka	6	6	8	7	7
8. mo	9	9	11	9	8
9. ke	14	14	15	10	12
10. ba	7	7	7	3	6
11. ga	13	13	14	12	13
12. fa	17	17	13	14	17
13. gore	15	19	16	29	16

14. di	10	10	10	16	11
15. wa	19	15	18	13	18
16. tsa	8	8	9	15	10
17. sa	18	22	22	21	22
18. se	16	16	17	17	15
19. re	35	38	21	11	20
20. tse	12	12	12	25	14
21. bo	25	27	27	35	25
22. kwa	20	28	20	19	21
23. mme	23	33	23	20	24
24. la	21	21	19	23	23
25. nna	26	26	32	32	27
26. ne	45	72	26	27	28
27. bone	42	60	28	36	30
28. tla	28	20	24	18	19
29. fela	46	47	33	49	39
30. jaaka	41	43	37	41	36
31. na	33	31	34	38	29
32. jwa	22	39	25	34	26
33. batho	24	36	39	50	33
34. jalo	52	67	40	54	44
35. yo	82	52	65	28	38
36. teng	37	66	42	42	45
37. tswa	53	78	53	78	50
38. dira	36	29	31	48	31
39. thata	63	98	52	60	62
40. bona	39	55	56	46	42
41. mongwe	64	24	55	44	32
42. pele	69	45	67	40	58
43. gagwe	- (290)	- (122)	91	26	41
44. bangwe	- (125)	- (147)	87	- (179)	- (121)
45. gape	64	93	57	100	64
46. ntse	- (102)	- (193)	80	75	85
47. neng	- (188)	- (136)	68	58	80
48. botswana	27	92	30	- (381)	47
49. yone	61	- (170)	44	- (250)	76
50. motho	77	63	- (112)	76	71
51. morago	- (101)	80	73	- (104)	90
52. itse	- (139)	- (351)	- (236)	88	- (145)
53. rona	- (128)	- (127)	64	43	46
54. nako	68	82	61	- (206)	69
55. ene	84	- (624)	79	53	84
56. gone	259	- (497)	- (143)	- (194)	- (249)
57. bua	- (345)	- (300)	- (311)	67	89
58. kana	- (498)	- (152)	- (102)	- (116)	- (130)
59. batla	- (143)	- (284)	- (152)	- (178)	- (134)
60. lefatshe	48	- (120)	50	57	48
61. tota	- (449)	- (434)	- (194)	- (417)	- (230)
62. tsaya	97	- (101)	94	- (113)	91
63. madi	80	91	29	- (107)	57
64. one	- (137)	- (290)	- (116)	- (265)	- (177)
65. tshwanetse	32	25	38	- (162)	35
66. bile	- (354)	85	- (101)	- (190)	- (146)

67. jaanong	- (327)	- (288)	- (186)	86	- (155)
68. jang	92	- (210)	- (163)	- (117)	- (112)
69. dilo	98	- (134)	- (134)	93	- (106)
70. nngwe	75	30	62	- (109)	54
71. dingwe	54	53	63	- (279)	66
72. rre	- (534)	- (2,527)	74	- (348)	- (227)
73. sentle	- (174)	- (196)	- (136)	- (270)	- (141)
74. jaana	- (634)	- (262)	- (214)	- (115)	- (135)
75. kgotsa	30	18	45	- (322)	34
76. sena	- (209)	- (370)	- (160)	95	- (158)
77. setse	- (332)	- (536)	- (161)	- (373)	- (243)
78. nne	90	- (114)	92	63	94
79. tiro	73	54	41	- (142)	63
80. leng	95	77	93	89	79
81. batswana	- (181)	- (738)	- (148)	- (1,682)	- (334)
82. supa	- (122)	- (232)	99	- (184)	- (125)
83. jo	74	89	60	74	51
84. gaborone	- (471)	- (1,110)	- (169)	- (10,298)	- (514)
85. sengwe	- (262)	- (112)	- (219)	- (133)	- (133)
86. ngwaga	- (121)	- (154)	46	- (188)	88
87. eo	- (182)	100	- (142)	- (221)	- (175)
88. bana	- (176)	- (335)	- (253)	- (123)	- (182)
89. tsone	94	- (213)	83	- (374)	- (157)
90. eng	- (237)	- (274)	- (277)	- (114)	- (140)
91. mokgosi	- (2,945)	- (3,850)	- (135)	- (844)	- (707)
92. tsotlhe	(113)	99	98	51	70
93. ntlha	62	94	- (123)	64	68
94. dumela	- (700)	- (339)	- (301)	- (543)	- (405)
95. tlhalosa	- (316)	- (549)	77	- (1,057)	- (162)
96. tsena	- (382)	- (201)	- (403)	- (138)	- (293)
97. gongwe	- (273)	- (329)	- (231)	81	- (215)
98. mangwe	- (123)	- (137)	82	- (443)	- (129)
99. gompieno	- (547)	- (612)	- (235)	- (425)	- (303)
100.seka	- (112)	- (303)	- (187)	- (668)	- (170)
Total	**64**	**59**	**71**	**61**	**68**

Table 67: Prose, Hansard, Call-in, Newspaper and PRONEWHANCAL

Whole corpus	Prose	Hansard	Call-in	Newspaper	PRONEWHANCAL
1. a	1	1	1	1	1
2. e	4	2	7	2	3
3. go	2	6	4	4	2
4. o	5	- (163)	9	3	10
5. le	3	- (108)	3	6	7
6. ya	10	13	12	8	12
7. ka	7	9	10	7	8
8. mo	9	11	11	10	11
9. ke	8	7	2	9	5
10. ba	6	3	5	5	4
11. ga	12	15	14	14	14
12. fa	11	10	13	12	13
13. gore	16	4	8	13	9

14. di	17	12	17	16	16
15. wa	18	42	20	18	21
16. tsa	21	20	30	17	23
17. sa	19	25	26	21	20
18. se	15	14	15	15	15
19. re	14	5	6	11	6
20. tse	25	18	21	22	22
21. bo	23	16	18	23	18
22. kwa	20	26	35	20	24
23. mme	22	17	19	24	19
24. la	29	44	34	25	29
25. nna	28	35	27	31	27
26. ne	13	21	24	19	17
27. bone	40	24	29	26	28
28. tla	24	39	23	28	25
29. fela	31	22	25	27	26
30. jaaka	37	52	54	35	39
31. na	33	37	28	34	30
32. jwa	41	- (154)	87	30	48
33. batho	43	29	44	32	32
34. jalo	48	69	59	36	45
35. yo	30	55	46	33	31
36. teng	46	38	36	41	40
37. tswa	45	93	63	52	66
38. dira	58	46	55	39	49
39. thata	55	- (104)	71	46	61
40. bona	27	61	43	47	37
41. mongwe	44	- (116)	58	40	57
42. pele	52	- (121)	88	56	70
43. gagwe	26	- (153)	- (104)	37	47
44. bangwe	97	87	82	48	84
45. gape	63	- (118)	75	63	75
46. ntse	32	45	39	53	33
47. neng	36	92	- (103)	42	51
48. botswana	- (799)	75	91	38	64
49. yone	- (103)	41	62	49	52
50. motho	38	- (110)	76	50	67
51. morago	47	- (199)	- (114)	57	88
52. itse	34	54	49	58	42
53. rona	83	33	33	45	36
54. nako	59	- (128)	72	61	77
55. ene	42	- (111)	64	43	72
56. gone	86	48	48	71	56
57. bua	49	32	45	72	34
58. kana	74	40	41	59	44
59. batla	61	100	61	88	86
60. lefatshe	- (185)	78	- (109)	51	90
61. tota	92	79	53	74	69
62. tsaya	73	74	69	81	68
63. madi	- (113)	57	70	44	60
64. one	- (140)	63	68	86	87
65. tshwanetse	100	73	81	64	76
66. bile	67	- (132)	90	69	83

67. jaanong	72	23	31	78	35
68. jang	82	- (115)	52	93	74
69. dilo	90	62	67	80	81
70. nngwe	93	- (134)	- (171)	85	- (120)
71. dingwe	- (131)	- (131)	- (139)	84	- (128)
72. rre	- (107)	- (215)	37	55	58
73. sentle	77	- (122)	79	94	82
74. jaana	84	53	57	95	55
75. kgotsa	70	- (551)	- (336)	65	- (123)
76. sena	- (108)	- (124)	- (127)	79	- (113)
77. setse	50	- (123)	- (111)	90	80
78. nne	- (137)	84	93	- (106)	99
79. tiro	- (114)	- (143)	- (138)	73	- (108)
80. leng	- (117)	28	32	- (111)	46
81. batswana	- (712)	58	77	67	62
82. supa	- (190)	- (213)	- (243)	89	- (151)
83. jo	- (109)	- (137)	- (121)	76	- (109)
84. gaborone	- (437)	- (332)	- (107)	- (101)	- (170)
85. sengwe	87	98	96	100	100
86. ngwaga	- (281)	- (332)	- (211)	70	- (143)
87. eo	- (126)	- (152)	- (113)	83	- (115)
88. bana	64	83	47	62	78
89. tsone	- (182)	68	- (108)	92	- (103)
90. eng	54	91	56	- (114)	59
91. mokgosi	- (916)	- (11,066)	- (1,346)	68	- (354)
92. tsotlhe	- (135)	- (252)	- (200)	- (144)	- (147)
93. ntlha	69	- (162)	- (178)	- (105)	- (111)
94. dumela	- (186)	- (305)	- (194)	- (186)	- (221)
95. tlhalosa	- (244)	- (394)	- (350)	96	- (204)
96. tsena	62	- (105)	97	- (126)	- (101)
97. gongwe	- (122)	64	74	- (102)	94
98. mangwe	- (290)	- (180)	- (203)	- (109)	- (191)
99. gompieno	- (192)	81	83	- (118)	- (102)
100. seka	- (525)	- (129)	86	- (103)	- (106)
Total	**74**	**65**	**78**	**88**	**80**

Below we summarise the results of Tables 65, 66, and 67 which show the similarities between the different text types and the whole corpus.

A

Poetry	62
Grammar	74
Chat-site	21
Plays	72
POEGRACHAPLA	59

B

Science	64
Politics	59
Business	71
Religious	61
SCIPOLBUSREL	68

C

Prose	74
Hansard	65
Call-in	78
Newspaper	88
PRONEWHANCAL	80

Tables 65, 66, and 67 reveal that the results of POEGRACHAPLA, SCIPOLBUSREL and PRONEWHANCAL depend on the text types that constitute them. Since these

three are made from samples taken from other text types they largely reflect the general trend found in such text types. The results are summarised in tables A, B and C above. Let us illustrate this phenomenon by looking at POEGRACHAPLA which comprises texts from Poetry, Grammar, Chat-site and Plays whose top 100 token-similarity with the top 100 tokens of the whole Setswana corpus are 62, 74, 21 and 72 respectively. POEGRACHAPLA has a token similarity of 59 with the whole corpus. Texts that make up POEGRACHAPLA in general have smaller similarity with the top 100 texts of the entire Setswana corpus. Consequently POEGRACHAPLA has little similarity with the top 100 texts of the entire Setswana corpus. These results compare well with those of group C which are on average higher. Prose, Hansard, Call-in and Newspaper's top 100 token-similarities with the top 100 tokens of the whole Setswana corpus are 74, 65, 78 and 88 respectively. Consequently PRONEWHANCAL has a higher corpus similarity of 80 with the most frequent 100 tokens of the whole Setswana corpus.

The averages for the four text types in A, B and C are 57, 64 and 76 respectively.

We therefore conclude that it is not enough to have a corpus with a variety of text types to generate large numbers of types. It is also crucial that the individual text types that comprise a corpus should individually have higher levels of types, as in Group C.

7.4.1 Comparison of the top 100 tokens of spoken and written Setswana

We conclude this section of experiments by comparing the most frequent 100 words of the spoken and the written part of the corpus to that of the most frequent 100 words of the entire corpus. The results follow in Table 68.

Table 68: Comparison of written and spoken components to the whole corpus

Whole corpus		**Written language**	**Spoken language**
1.	a	1	1
2.	e	3	2
3.	go	2	4
4.	o	5	17
5.	le	4	16
6.	ya	8	12
7.	ka	7	8
8.	mo	10	11
9.	ke	9	6
10.	ba	6	3

11.	ga	11	13
12.	fa	12	10
13.	gore	16	7
14.	di	15	14
15.	wa	18	29
16.	tsa	19	13
17.	sa	20	25
18.	se	14	15
19.	re	13	5
20.	tse	23	20
21.	bo	24	18
22.	kwa	22	28
23.	mme	25	19
24.	la	27	42
25.	nna	28	31
26.	ne	17	23
27.	bone	35	26
28.	tla	26	32
29.	fela	31	24
30.	jaaka	36	58
31.	na	33	33
32.	jwa	34	- (136)
33.	batho	37	35
34.	jalo	48	62
35.	yo	29	49
36.	teng	45	38
37.	tswa	50	82
38.	dira	47	53
39.	thata	53	99
40.	bona	32	55
41.	mongwe	43	95
42.	pele	54	- (113)
43.	gagwe	30	- (133)
44.	bangwe	77	91
45.	gape	67	- (111)
46.	ntse	40	47
47.	neng	44	98
48.	botswana	99	81
49.	yone	82	45
50.	motho	41	- (103)
51.	morago	58	- (173)
52.	itse	39	52
53.	rona	57	36
54.	nako	63	- (110)
55.	ene	49	89
56.	gone	95	50
57.	bua	55	37
58.	kana	66	44
59.	batla	69	90
60.	lefatshe	96	85
61.	tota	- (102)	70
62.	tsaya	85	67
63.	madi	73	63
64.	one	- (109)	65
65.	tshwanetse	62	75
66.	bile	76	- (126)
67.	jaanong	78	27
68.	jang	90	93
69.	dilo	89	66
70.	nngwe	70	- (144)
71.	dingwe	- (111)	- (137)
72.	rre	- (108)	- (108)
73.	sentle	93	- (114)
74.	jaana	98	56
75.	kgotsa	52	- (394)
76.	sena	- (125)	- (134)
77.	setse	74	- (131)
78.	nne	- (121)	88
79.	tiro	86	- (143)
80.	leng	- (115)	34
81.	Batswana	- (179)	64
82.	supa	- (143)	- (190)
83.	jo	88	- (147)
84.	Gaborone	- (253)	- (232)
85.	sengwe	92	97
86.	ngwaga	- (159)	- (278)
87.	eo	- (134)	- (132)
88.	bana	64	72
89.	tsone	- (158)	74
90.	eng	61	71
91.	mokgosi	- (281)	- (3,863)
92.	tsotlhe	- (120)	- (221)
93.	ntlha	75	- (174)
94.	dumela	- (192)	- (262)
95.	tlhalosa	- (185)	- (342)
96.	tsena	87	- (106)
97.	gongwe	- (119)	68
98.	mangwe	- (213)	- (185)
99.	gompieno	- (191)	83
100.	seka	- (268)	- (125)
Total		**81**	**71**

Ninety four percent of the entire Setswana corpus is written language and only 6% is spoken language component. The effects of this phenomenon are reflected in the results. Eighty one of the top 100 words of the written component of the corpus are

found in the most frequent 100 words of the whole corpus. On the other hand, 71 words of the top 100 words of the spoken component are found amongst the most frequent 100 words in the entire corpus. The written component of the corpus is much more diverse in terms of the kind of texts it comprises while comparatively the spoken component is limited. This may explain the differences between the two.

7.4.2 Comparison of the top 100 tokens of spoken and written parts of the BNC

Below we compare our results with those of the BNC to determine the quality of our results in comparison to those of a larger balanced English corpus. The BNC lists in Tables 69-72 are from Kilgarriff's website (www.kilgarriff.co.uk). We start off by first listing the top 100 words of the whole BNC, and those of the written and spoken components. Table 69 lists the most frequent 100 words of the BNC. Table 70 comprises the most frequent 100 words of the written subcorpus. Table 71 contains the most frequent words of the context governed section of the spoken subcorpus. Table 72 gives the BNC's top 100 words of the demographic section of the spoken corpus (See Appendix 7 for the BNC POS codes).

Table 69: The BNC top 100 words of the whole corpus

Freq	**word**	**POS**
6187267	the	at0
2941444	of	Prf
2682863	and	Cjc
2126369	a	at0
1812609	in	prp
1620850	to	to0
1089186	it	pnp
998389	is	vbz
923948	was	vbd
917579	to	prp
884599	i	pnp
833360	for	prp
695498	you	Pnp
681255	he	Pnp
662516	be	Vbi
652027	with	Prp
647344	on	Prp
628999	that	Cjt
507317	by	Prp
478162	at	Prp
470943	are	Vbb
462486	not	xx0
461945	this	dt0
454096	but	Cjc
442545	's	Pos
433441	they	Pnp
426896	his	Dps
413532	from	Prp
409012	had	Vhd
380257	she	Pnp
372031	which	Dtq
370808	or	Cjc
358039	we	Pnp
343063	an	at0
332839	n't	xx0
325048	's	Vbz
322824	were	Vbd
286913	that	dt0
268723	been	Vbn
268490	have	Vhb
260919	their	Dps
259431	has	Vhz
255188	would	Vm0
249466	what	dtq
244822	will	vm0
239460	there	ex0
237089	if	cjs
234386	can	vm0
227737	all	dt0
218258	her	dps
208623	as	cjs
205432	who	pnq
205195	have	vhi
196635	do	vdb
194800	that	Cjt-dt0
190499	one	Crd
185277	said	Vvd
173414	them	Pnp

171174	some	dt0
168387	could	vm0
165014	him	pnp
163469	into	prp
163081	its	dps
160652	then	av0
156111	two	crd
155417	when	avq-cjs
154288	up	avp
153679	time	nn1
152619	my	dps
150958	out	avp
147324	so	av0
143405	did	vdd

142118	about	prp
138334	your	dps
137801	now	av0
137151	me	pnp
137026	no	at0
134029	more	av0
129451	other	aj0
125465	just	av0
125442	these	dt0
124884	also	av0
123916	people	nn0
123655	any	dt0
118699	first	ord
115994	only	av0
114655	new	aj0

113024	may	vm0
111538	very	av0
111236	should	vm0
111083	as	cjs-prp
108988	like	prp
108710	her	pnp
108618	than	cjs
106427	as	prp
101508	how	avq
96080	well	av0
95313	way	nn1
95001	our	dps
91583	as	av0

Table 70: The BNC top 100 words of the written corpus component

Freq	**word**	**POS**
5776384	the	at0
2789403	of	prf
2421302	and	cjc
1939617	a	at0
1695860	in	prp
1468146	to	to0
892937	is	vbz
845350	to	prp
839964	was	vbd
834957	it	pnp
768898	for	prp
606027	with	prp
605749	he	pnp
603178	be	vbi
590305	on	prp
580267	i	pnp
561041	that	cjt
490673	by	prp
435574	at	prp
426207	you	pnp
425898	's	pos
422562	are	vbb
413798	not	xx0
413737	his	dps
404140	this	dt0
390876	from	prp
389108	but	cjc
386510	had	vhd
349120	which	dtq
337345	she	pnp
336599	they	pnp
335976	or	cjc

323963	an	at0
294301	were	vbd
249636	we	pnp
247596	their	dps
247131	been	vbn
242854	has	vhz
225582	have	vhb
225381	will	vm0
221172	would	vm0
211159	her	dps
206150	n't	xx0
201616	there	ex0
197483	can	vm0
195515	all	dt0
193757	as	cjs
189926	if	cjs
186984	who	pnq
173582	what	dtq
170417	have	vhi
165805	that	cjt-dt0
161742	that	dt0
160935	said	vvd
159578	its	dps
157972	one	crd
157300	's	vbz
152395	into	prp
151029	him	pnp
150609	some	dt0
148165	could	vm0
140989	them	pnp
138167	when	avq-cjs

134941	time	nn1
129196	out	avp
128980	my	dps
127987	two	crd
127430	up	avp
124543	no	at0
124501	then	av0
123686	more	av0
123315	do	vdb
119113	also	av0
116367	other	aj0
115946	about	prp
112278	these	dt0
110963	me	pnp
108775	first	ord
108669	your	dps
108593	so	av0
108462	did	vdd
108392	new	aj0
108301	now	av0
108088	may	vm0
108043	any	dt0
105560	as	cjs-prp
105411	only	av0
102554	as	prp
102516	people	nn0
101495	than	cjs
100822	her	pnp
99069	should	vm0
87862	like	prp
87705	as	av0
87034	between	prp

86823	very	av0
85826	just	av0

82920	many	dt0
82878	years	nn2

82343	way	nn1

Table 71: The BNC top 100 words of the context-governed spoken corpus

Freq	**word**	**POS**
295636	the	at0
170675	and	cjc
136692	i	pnp
134074	you	pnp
126064	it	pnp
117906	a	at0
117140	of	prf
105313	to	to0
82272	in	prp
75509	's	vbz
75237	we	pnp
70296	is	vbz
67160	er	unc
62789	that	dt0
54810	that	cjt
52865	they	pnp
49147	was	vbd
49078	n't	xx0
48932	to	prp
45807	erm	unc
41895	for	prp
40640	be	vbi
38220	this	dt0
37755	but	cjc
37369	what	dtq
35798	on	prp
31824	are	vbb
31079	do	vdb
29795	if	cjs
29442	with	prp
28845	at	prp
27288	not	xx0
27184	he	pnp

25366	've	vhb
25099	have	vhb
24926	there	ex0
24418	're	vbb
23995	would	vm0
23049	yeah	itj
23033	or	cjc
22778	so	av0
22382	well	av0
21947	yes	itj
21606	can	vm0
21524	that	cjt-dt0
21072	one	crd
20076	just	av0
19464	which	dtq
19348	think	vvb
18841	know	vvb
18541	then	av0
18486	have	vhi
18220	very	av0
17961	all	dt0
17953	were	vbd
17915	now	av0
17734	two	crd
17403	about	prp
17089	from	prp
16711	people	nn0
16679	them	pnp
16116	got	vvn
16107	there	av0
15188	your	dps
14854	because	cjs
14696	did	vdd
14293	do	vdi

14268	an	at0
14247	will	vm0
14171	been	vbn
13500	by	prp
12929	had	vhd
12844	right	av0
12753	some	dt0
12641	up	avp
12596	'll	vm0
12122	could	vm0
12068	going	vvg
12012	'm	vbb
11986	who	pnq
11950	has	vhz
11851	no	itj
11595	my	dps
11569	time	nn1
11541	three	crd
11387	as	cjs
11202	out	avp
11184	how	avq
10891	mm	itj
10821	me	pnp
10809	mean	vvb
10758	oh	itj
10692	get	vvi
10589	their	dps
10431	any	dt0
10195	our	dps
10114	so	cjs
9964	's	pos
9547	when	avq-cjs
9522	actually	av0

Table 72: The BNC top 100 words of the demographic spoken corpus

Freq	**word**	**POS**
167640	i	pnp
135217	you	pnp
128165	it	pnp
115247	the	at0
92239	's	vbz
90886	and	cjc
77611	n't	xx0
68846	a	at0

62382	that	dt0
58810	yeah	itj
48322	he	pnp
47391	to	to0
43977	they	pnp
42241	do	vdb
41654	oh	itj
38515	what	dtq
35156	is	vbz

34901	of	prf
34837	was	vbd
34477	in	prp
33763	she	pnp
33166	we	pnp
31662	no	itj
30177	well	av0
27233	but	cjc
23297	to	prp

22779	've	vhb
22567	for	prp
22016	got	vvn
21907	mm	itj
21586	know	vvb
21400	not	xx0
21370	er	unc
21241	on	prp
20247	did	vdd
19720	're	vbb
19585	this	dt0
19563	just	av0
19212	'll	vm0
18698	be	vbi
18284	there	av0
18051	said	vvd
17898	yes	itj
17809	have	vhb
17610	then	av0
17368	if	cjs
16619	erm	unc
16558	with	prp
16557	are	vbb
16292	have	vhi
15953	so	av0

15746	them	pnp
15367	me	pnp
15297	can	vm0
14477	your	dps
14261	all	dt0
14217	up	avp
14048	'm	vbb
13743	at	prp
13148	that	cjt
12918	there	ex0
12539	get	vvi
12044	my	dps
11952	like	prp
11911	do	vdi
11799	or	cjc
11585	now	av0
11455	one	Crd
11378	cos	cjs
10570	were	vbd
10560	out	avp
10541	think	vvb
10484	go	vvi
10468	mean	vvb
10390	two	crd
10251	going	vvg

10120	know	vvi
10051	na	to0
10021	would	vm0
9573	had	vhd
9163	really	av0
9161	right	av0
8984	one	pni
8896	him	pnp
8812	's	vhz
8769	about	prp
8443	here	av0
8367	how	avq
8100	could	vm0
8087	ca	vm0
8085	gon	vvg
7812	some	dt0
7807	does	vdz
7703	when	avq-cjs
7545	good	aj0
7471	that	cjt-dt0
7461	on	avp
7421	been	vbn
7371	go	vvb
7344	down	avp

The results of Table 73 below are from Leech et al. (2001: 144) since Kilgarriff's website does not have readily available lists for the spoken component of the BNC.

Table 73: The BNC top 100 words of the spoken part of the whole corpus

39605	The	Det
29448	I	Pron
25957	you	Pron
25210	and	Conj
24508	It	Pron
18637	A	Det
17677	's	Verb
14912	to	Inf
14550	of	Prep
14252	that	DetP
12212	n't	Neg
11609	in	Prep
10448	we	Pron
10164	is	Verb
9594	do	Verb
9333	they	Pron
8542	er	Uncl
8097	was	Verb
7890	yeah	Int
7488	have	Verb
7313	what	DetP

7277	he	Pron
7246	that	Conj
6950	to	Prep
6366	but	Conj
6239	for	Prep
6029	erm	Uncl
5790	be	Verb
5659	on	Prep
5627	this	DetP
5550	know	Verb
5310	well	Adv
5067	so	Adv
5052	oh	Int
5025	got	Verb
4735	've	Verb
4693	not	Neg
4663	are	Verb
4544	if	Conj
4446	with	Prep
4388	no	Int
4255	're	Verb

4136	she	Pron
4115	at	Prep
4067	there	Ex
3977	think	Verb
3840	yes	Int
3820	just	Adv
3644	all	DetP
3588	can	VMod
3474	then	Adv
3464	get	Verb
3368	did	Verb
3357	or	Conj
3278	would	VMod
3163	mm	Int
3126	them	Pron
3066	'll	VMod
3034	one	Num
2894	there	Adv
2891	up	Adv
2885	go	Verb
2864	now	Adv

2859	your	Det
2835	had	Verb
2749	were	Verb
2730	about	Prep
2710	two	Num
2685	said	Verb
2532	one	Pron
2512	'm	Verb
2507	see	Verb
2444	me	Pron
2373	very	Adv
2316	out	Adv
2278	my	Det

2255	when	Conj
2250	mean	Verb
2209	right	Adv
2208	which	DetP
2178	from	Prep
2174	going*	Verb
2116	say	Verb
2082	been	Verb
2063	people	NoC
2039	because*	Conj
1986	some	DetP
1949	could	VMod
1890	will	VMod

1888	how	Adv
1849	on	Adv
1846	an	Det
1819	time	NoC
1780	who	Pron
1776	want	Verb
1762	like	Prep
1737	come	Verb
1727	really	Adv
1721	three	Num
1663	by	Prep

Below we compare the top 100 words of the whole corpus (WC*) against the written (WR*) and spoken (SP*) subcorpora of the BNC against the top 100 words of the whole corpus. The results of this comparison are subsequently compared to the results of experiments carried on the Setswana corpus above.

Table 74: Comparison of the top 100 words of the BNC against the top 100 words of the written and spoken subcorpora

N	WC	WR	SP
1	the	1	1
2	of	2	9
3	and	3	4
4	a	4	6
5	in	5	12
6	to	6	-
7	it	10	5
8	is	7	14
9	was	9	18
10	to	8	8
11	i	16	2
12	for	11	26
13	you	20	3
14	he	13	22
15	be	14	28
16	with	12	40
17	on	15	91
18	that	17	23
19	by	18	100
20	at	19	44
21	are	22	38
22	not	23	37
23	this	25	30
24	but	27	25
25	's	21	-
26	they	31	16

27	his	24	-
28	from	26	81
29	had	28	65
30	she	30	43
31	which	29	80
32	or	32	54
33	we	35	13
34	an	33	92
35	n't	43	11
36	's	57	7
37	were	34	66
38	that	53	10
39	been	37	84
40	have	39	20
41	their	36	-
42	has	38	-
43	would	41	55
44	what	50	21
45	will	40	89
46	there	44	45
47	if	48	39
48	can	45	50
49	all	46	49
50	her	42	-
51	as	47	-
52	who	49	94
53	have	51	-

54	do	72	15
55	that	52	23
56	one	56	59
57	said	54	69
58	them	62	57
59	some	60	87
60	could	61	88
61	him	59	-
62	into	58	-
63	its	55	-
64	then	70	51
65	two	67	68
66	when	63	77
67	up	68	61
68	time	64	93
69	my	66	76
70	out	65	75
71	so	80	33
72	did	81	53
73	about	75	67
74	your	74	64
75	now	83	63
76	me	77	73
77	no	69	41
78	more	71	-
79	other	74	-
80	just	97	48

81	these	76	-
82	also	73	-
83	people	89	85
84	any	85	-
85	first	78	-
86	only	87	-
87	new	82	-
88	may	84	-

89	very	96	74
90	should	92	-
91	as	86	-
92	like	93	96
93	her	91	-
94	than	90	-
95	as	88	-
96	how	-	90

97	well	-	32
98	way	100	-
99	our	-	-
100	as	94	-
TOTAL		**97**	**72**

N = Word rank

WC = Whole corpus

WR = Written subcorpus

SP = Spoken subcorpus

Ninety percent of the BNC is written material while 10% is transcribed speech. Ninety seven of the top 100 words of the written component of the corpus are found in the most frequent 100 words of the whole corpus. On the other hand, only 72 words of the top 100 words of the spoken component are found amongst the most frequent 100 words of the entire corpus. Table 75 below shows a comparison of the results of the BNC and of those of the Setswana corpus.

Table 75: Comparison of BNC and Setswana

Corpus	Written Component	Spoken component
BNC	97	72
Setswana Corpus	81	71

When we compare the BNC results with those of the Setswana corpus components we find that 71 of the Setswana spoken subcorpus' most frequent 100 words are found amongst the most frequent 100 Setswana words. Eighty one of most frequent 100 written words are found amongst the most frequent 100 words of the complete Setswana corpus. The results are fairly similar, particularly those of the spoken part of the corpus. The corpus components are also comparable since the BNC has 90% written material and 10% transcribed speech while Setswana corpus is 94% written material and 6% transcribed speech.

Both the top 100 written and spoken components of the corpus do not have all the words found in the top 100 words of the whole corpus. It is however worth noting that

the written and spoken components are complimentary since words which are absent in one subcorpus may be found in another subcorpus.

7.5 A direct comparison of Setswana spoken and written corpus components

Having compared the most frequent 100 words of spoken and written language against the most frequent 100 words of the corpus by seeing which words of each subcorpus are present in the top 100 words of the entire corpus, we now turn to comparing both subcorpus components directly with each other. We use Wordsmith Tools' wordlist program to compare the wordlists directly. This program is exactly the same as the keywords program discussed previously and uses log likelihood statistic as well.

> The procedure compares all the words in both lists and reports on all those which appear significantly more often in one than the other, including those which appear more than a minimum number of times in one even if they do not appear at all in the other (Scott, 2004-2006: 106).

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent to spoken language. At the end of the listing are those words which are outstandingly infrequent in spoken language but are key to written language. In Table 76, we give the most frequent 30 words in spoken language and Table 77 gives the most infrequent words in spoken language, or the most key words in the written language.

Table 76: Outstandingly frequent spoken language

N	Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
1	gore	26,232	3.12	98,272	0.81	28,857.99
2	re	27,313	3.24	130,324	1.08	21,884.16
3	mr	2,625	0.31	515		11,627.94
4	ko	3,528	0.42	3,969	0.03	9,469.37
5	ke	27,070	3.22	214,179	1.77	7,551.83
6	rraetsho	1,592	0.19	257		7,251.76
7	hansard	1,191	0.14	0		6,514.12

8	honourable	1,126	0.13	4		6,105.93
9	jaanong	3,722	0.44	9,556	0.08	5,892.05
10	leng	3,301	0.39	7,355	0.06	5,856.40
11	speaker	1,071	0.13	8		5,764.31
12	the	4,960	0.59	19,603	0.16	5,055.83
13	motsamaisa	991	0.12	85		4,836.79
14	2002	800	0.10	0		4,375.22
15	member	820	0.10	31		4,222.55
16	resumed	756	0.09	0		4,134.54
17	debate	799	0.09	46		4,018.69
18	page	1,331	0.16	1,164		3,988.55
19	dipuisanyo	980	0.12	343		3,891.60
20	ee	2,104	0.25	4,489	0.04	3,854.29
21	palamente	1,124	0.13	896		3,493.26
22	motion	654	0.08	11		3,466.05
23	ba	28,570	3.39	283,076	2.34	3,364.43
24	bill	641	0.08	51		3,148.24
25	of	2,812	0.33	11,011	0.09	2,895.52
26	rona	3,101	0.37	13,185	0.11	2,876.46
27	bua	3,090	0.37	13,985	0.12	2,631.98
28	ra	1,839	0.22	5,575	0.05	2,500.78
29	yone	2,353	0.28	9,187	0.08	2,430.83
30	kana	2,581	0.31	10,927	0.09	2,406.97

Table 76 and 77 are sorted on the basis of keyness or log likelihood statistic listed on the seventh column on the extreme right of the table. RC. Freq. and RC. % refers to the word frequency of the reference corpus and reference corpus's word percentage respectively.

Table 77: Outstandingly infrequent spoken tokens

N	Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
1,954	Yo	2,086	0.25	34,334	0.28	-173,359.36
1,955	gagwe	728	0.09	33,421	0.28	-177,059.83
1,956	nna	3,433	0.41	39,498	0.33	-194,254.70
1,957	la	2,669	0.32	45,661	0.38	-231,733.53
1,958	mme	6,936	0.82	52,649	0.43	-249,051.45
1,959	bo	7,160	0.85	54,427	0.45	-257,615.50
1,960	tla	3,356	0.40	51,181	0.42	-258,048.47
1,961	tse	6,061	0.72	62,008	0.51	-303,327.06
1,962	kwa	3,646	0.43	64,275	0.53	-328,216.78
1,963	i	2,878	0.34	68,621	0.57	-356,843.38
1,964	sa	3,881	0.46	76,545	0.63	-394,662.47
1,965	gore	26,232	3.12	98,272	0.81	-423,485.22
1,966	tsa	4,523	0.54	86,900	0.72	-448,427.16
1,967	wa	3,564	0.42	90,486	0.75	-474,360.16
1,968	ne	4,364	0.52	92,154	0.76	-478,703.91
1,969	di	9,586	1.14	113,319	0.94	-568,445.81
1,970	re	27,313	3.24	130,324	1.08	-590,392.75

1,971	se	9,445	1.12	122,154	1.01	-618,585.81
1,972	fa	11,418	1.36	130,440	1.08	-655,166.81
1,973	ga	10,334	1.23	138,306	1.14	-705,009.69
1,974	mo	11,156	1.33	180,148	1.49	-940,332.00
1,975	ke	27,070	3.22	214,179	1.77	-1,060,745.50
1,976	ya	10,538	1.25	215,238	1.78	-1,149,694.00
1,977	ka	19,592	2.33	268,149	2.21	-1,415,689.88
1,978	ba	28,570	3.39	283,076	2.34	-1,461,408.50
1,979	o	7,328	0.87	320,525	2.65	-1,816,507.13
1,980	le	8,687	1.03	345,885	2.85	-1,968,098.25
1,981	e	30,954	3.68	372,429	3.07	-2,006,141.63
1,982	go	27,937	3.32	385,650	3.18	-2,107,078.75
1,983	a	33,154	3.94	643,503	5.31	-3,936,417.75

A look at Table 76 results shows a high level of parliament terminology as evidenced by the following, *Rraetsho* (Sir) (6), Hansard (7), Honourable (8), Speaker (11), *Motsamaisa Dipuisanyo* (13, 19) (Speaker), Member (15), Resumed (16) and Debate (17). This is expected since spoken text in the corpus is dominated by parliamentary Hansard documents.

"*Gore*" (that) is the most key which scores 28,857.99 on the keyness column and "*A*" is the most outstandingly infrequent with -3,936,417.75 on the keyness column. The top infrequent words include mostly words which are members of the class of closed words such as *Ka* (with) (1,977), *Ya* (of) (1,976), *Go* (1,982) (to), *E* (1,981) (it), *Le* (1,980) (and), *O* (1,979) (he/she), *Ba* (1,978) (they, those), *Gore* (1965) (that), *Gagwe* (1,955) (his/hers). We would expect most of these words to appear high in the spoken subcorpus however an inspection of the most frequent words in the written corpus in Table 68 shows that these words rank high in the written corpus as well. Two matters may be responsible for their showing in the written subcorpus: first, is the size of the written subcorpus which is large compared to the spoken part of the corpus. The written subcorpus is 94% of the whole corpus while the spoken subcorpus is only 6%. Second, the 6% of the spoken subcorpus has a large Hansard section. Hansard material occupies 73% of the whole spoken subcorpus.

The comparison is significant since it reveals that there are distinctions between spoken and written language. This observation is relevant to corpus design since recognition of the distinction of spoken and written and spoken language should influence corpus compilers to sample both written and spoken language for inclusion

in a corpus.

7.6 Comparison of opportunistic and balanced corpora

In the past few experiments we have investigated different text types and through keyword analysis isolated those words which are particular to them. The experiments were intended to test whether different text types contribute distinct words. These findings are relevant to corpus design for lexicography in general, and particularly to this thesis. The recognition that different text types contribute different numbers of words should influence lexicographers compiling dictionaries on the basis of corpus evidence to pay closer attention to corpus design to ensure the broadest coverage possible of text types in a corpus. This is for the reasons that the quality of retrieved information for lexicographic purposes depends on the information input at the stage of corpus construction.

To further test whether text type diversity is crucial to the words selected for inclusion in a dictionary, we compare two 5,000-word list chunks. The first chunk simulates an opportunistic corpus (also called convenience sample (Borin, 2000: 76)) with its text type limitations since it is derived exclusively from prose text. We use the prose text since many readily available text materials in many African languages is of a prose type. The majority of such text would comprise novels.

While in many African languages most readily available text will be prose, in other contexts such text may be newspaper text or web text (see Borin, 2000 and Mair, 1992). For instance 900 million words of Afrikaans texts in the Media24[27] archives could be used as corpus material. Such corpora would be heavily skewed towards a single text type and may not be taken to represent language variability that exists in a speech community. A good illustration of this is MacLeod and Grishman (2000) who report on the creation of two machine readable dictionaries COMLEX Syntax and NOMLEX produced at New York University, in which they used the BNC and the Brown Corpus to which they added a large amount of newspaper text. COMLEX

[27] www.media24.com

contained 7 MB of the Brown Corpus, 27 MB of Wall Street Journal, 30 MB of San Jose Mercury, 29.5 MB of Associate Press text and 1.5 MB miscellaneous selections from the Treebank Literature. They illustrate how an increase in the Brown Corpus (which is generally regarded as balanced) of 1,329% (thus more than thirteen times) resulted in a skewed or inadequate corpus:

> First of all, the make-up of the POS corpus, with its preponderance of newspaper text, skewed the choice of high-frequency verbs. This can be seen by comparing the frequency-ranked list from this corpus with that from Brown, a more balanced corpus. Among the top 50 verbs from our corpus, quite a few (business-related) verbs were not in the top 50 from Brown, including *sell*, *rise*, *buy*, *pay*, and *increase*. In fact, some were not even in the top 750 from Brown, such as *post*, *boost*, *invest*, *value*, and *resign* (MacLeod and Grishman, 2000: 142).

Their results show that media publications such as texts from newspapers and journals mostly available in large quantities if used indiscriminately can skew a corpus (see also Čermák and Křen, 2005). Their experiment therefore offer support to the position that the opportunistic approach to corpus building runs the risk of creating a skewed corpus that does not adequately capture the linguistic rich diversity of a language.

Other researchers have also argued against an opportunistic corpus compilation approach. For instance Biber argues for corpus diversity by pointing that:

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

While we share Biber's position on corpus composition, his argument needs to be tested. To test the text type variability assumption the most frequent 5,000 words were

derived from the prose text and compared with 5,000 words from a variety of text types. The second wordlist of 5,000 words mirrors a balanced corpus, while the first wordlist mirrors an opportunistic one. It was compiled with 500 top keywords from the following 10 text types:

I. Newspaper Text
II. Religious Text
III. Chat-site Text
IV. Hansard Text
V. Poetry Text
VI. Prose text
VII. Politics Text
VIII. Science Text
IX. Call-in Text
X. Business Text

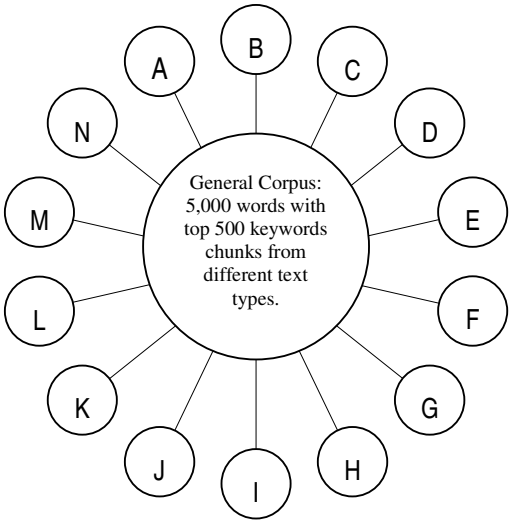
The purpose of comparing the two 5,000-word lists should be by now apparent. It is to measure which of the two lists covers a broad scope of linguistic varieties similar to the one found in the range of varieties of Setswana language use. While we acknowledge that both 5,000-word lists could be used in the compilation of dictionaries, we do however want to measure for wide linguistic coverage in both lists. The question we want to answer is whether the diversity of text types in corpus compilation adds significant value to the quality of dictionary entries by contributing broad word coverage or whether broad word coverage may be attained from a corpus compiled from a single text type, such as prose text.

However, the concept of broad text coverage should not be perceived as restricted to the compilation of corpora for general dictionaries. Even corpora for dictionaries of specialised areas like science and linguistics should demonstrate broad text coverage. This is because specialised areas tend to comprise a variety of even more specialised minute areas. For instance, the area of science is broad; it covers physics, chemistry, biology, engineering, physiology and a variety of other science subjects. The area of linguistics is equally broad comprising morphology, phonology, syntax, semantics, sociolinguistics, psycholinguistics, lexicology, lexicography, computational linguistics and a variety of other areas of language study. Corpora for dictionaries of specialised areas such as linguistics and science should (just like corpora for general dictionaries are compiled with a broad coverage of text types of the general language) also be compiled using a broad range of the text types that constitute the specialised area.

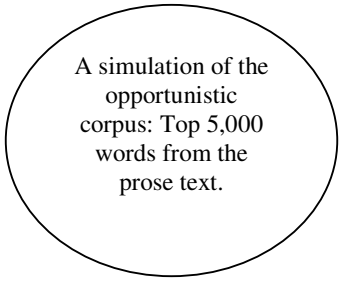
We graphically illustrate how the two 5,000 wordlists are compiled. On one hand is 500-word chunks from different sources compiled together to form a 5,000 wordlist and, on the other hand is the most frequent 5,000 words from a single text type, prose text. From henceforth we will refer to the 5,000 words derived from diverse sources as a combined list.

Figure 17: 5,000 words from a variety of sources

5,000 words from diverse sources



5,000 words from a single source



The 5,000 words from diverse sources were compiled by first sampling the top 500 tokens from each of text types. Each sampled token was sampled with its frequency from its text type. This resulted with overlaps and amongst overlapping tokens, tokens with lower frequencies were deleted and the one with a higher or the highest frequency was kept. We then added 50-tokens incrementally from each of the ten text types, deleting any overlaps until we got 5,000 tokens. Any extra tokens after reaching the 5,000-token were deleted. The tokens were ordered on the basis of their frequencies from their text types. Besides the ordering of tokens within the 5,000 tokens from diverse text types, the frequencies are not used to make comparisons between the two 5,000-token wordlists.

In Table 78 we give the results of the top 100 words from both lists.

Table 78: Top 100 tokens of Prose and Combined list

no	Prose list	Combined list
1.	a	go
2.	go	le
3.	le	ba
4.	e	ka
5.	o	ke
6.	ba	mo
7.	ka	fa
8.	ke	ga
9.	mo	ne
10.	ya	se
11.	fa	i
12.	ga	wa
13.	ne	sa
14.	re	o
15.	se	e
16.	gore	kwa
17.	di	ya
18.	wa	re
19.	sa	gore
20.	kwa	mme
21.	tsa	tla
22.	mme	gagwe
23.	bo	ie
24.	tla	bona
25.	tse	nna
26.	gagwe	the
27.	bona	to
28.	nna	yo
29.	la	fela
30.	yo	posted
31.	fela	ntse
32.	ntse	na
33.	na	itse
34.	itse	tsa
35.	i	neng
36.	neng	and
37.	jaaka	you
38.	motho	motho
39.	lo	di
40.	bone	ene
41.	jwa	on
42.	ene	of
43.	batho	bone
44.	mongwe	at
45.	tswa	by
46.	teng	is
47.	morago	tswa
48.	jalo	that
49.	bua	bo
50.	setse	morago
51.	rile	setse
52.	pele	kwa
53.	monna	tse
54.	eng	rile
55.	thata	pele
56.	ngwana	monna
57.	utlwa	eng
58.	dira	in
59.	nako	ngwana
60.	me	utlwa
61.	batla	are
62.	tsena	nako
63.	gape	it
64.	bana	morena
65.	pelo	batla
66.	letsatsi	tsena
67.	bile	gape
68.	gonne	la
69.	ntlha	pelo
70.	kgotsa	lo
71.	tsamaya	letsatsi
72.	jaanong	bile
73.	tsaya	gonne
74.	kana	gagwe
75.	mosadi	ntlha
76.	gago	tsamaya
77.	sentle	mosadi
78.	wena	sentle
79.	tle	tle
80.	kgosi	for
81.	kae	this
82.	jang	kae
83.	rona	matlho
84.	jaana	not
85.	matlho	botswana
86.	gone	we
87.	sengwe	ena
88.	ena	have
89.	ie	jaanong
90.	dilo	kete
91.	kete	modimo
92.	tota	ise
93.	nngwe	godimo
94.	ise	rata
95.	godimo	gago
96.	rata	twe
97.	bangwe	ko

98.	twe	ja
99.	ja	be

100.	tshwanetse	sepe

At the top of both the prose and combined frequency wordlist are the expected functional words which normally occupy the top rank of frequency lists in a variety of text types. These amongst other words include *a, go, e, le, o, ba, ka, ke, mo, fa, ya, ga, ne, mo* and *se*. The top words are therefore fairly similar to those found in the highest frequency position of the entire corpus. The most frequent words of both lists are therefore not very different from each other save minor differences of various words being at different positions of rank which are not very far from each other. This is positive for both lists since it means that both lists in general capture the most frequent words in the language. For lexicography, it means that if dictionaries were compiled using the two lists, the most frequent words, which in many instances are functional words, would not be excluded from the dictionary.

However to see if the different lists offer significant differences we must inspect the different parts of the two lists preferably looking for the inclusion of words from different text types. We will therefore discuss the inclusion of words in the different lists.

We start looking at religious text. We consider those words which characterise Christianity and traditional Setswana beliefs (TSB). We choose these two since they are followed by the largest percentage of the population with Christianity estimated at 68% and TSB at 30% (Humphries, 2003: 166). We focus on the following words:

Table 79: Christian terms

Setswana terms	English
1. *Jeso*	Jesus
2. *Keresete*	Christ
3. *Modimo*	God
4. *Baebele*	Bible
5. *Bakeresete*	Christians
6. *Legodimo*	Heaven

Table 80: TSB terms

TSB terms	English
1. *Badimo*	ancestors
2. *Moloi*	witch/wizard
3. *Setlhabelo*	sacrifice
4. *Dipheko*	charms
5. *Ditaola*	divination bones
6. *Matwetwe*	traditional doctor

In Tables 81 and Table 82 we offer the results pf the comparisons by showing the rank the words occupy in Prose and combined lists. A discussion of the results follows their presentation.

Table 81: Christian terms and their ranks on the two lists

Setswana	English	Prose text	Combined text
Jeso	Jesus	-	1,222
Baebele	Bible	-	1,303
Bakeresete	Christians	-	2,698
Legodimo	heaven	1,340	1,065
Keresete	Christ	4,855	678
Modimo	God	219	91

Table 82: TSB terms and their ranks on the two lists

Setswana	English	Prose text	Combined text
Badimo	Ancestors/gods	360	-
Moloi	Witch/wizard	1,701	3,319
Setlhabelo	Sacrifice	-	834
Dipheko	Charms	1,944	1731
Ditaola	Divination bones	3,277	2,756
Matwetwe	Traditional doctor	-	3,012

The constant result in both tables is that the combined text numbers are ranked higher in the list compared to the prose words save for *moloi* (witch/wizard) which appears higher in prose text. Some of the gaps between words in the two lists are significantly higher. For instance, the difference in *Keresete* (Christ) is at 4177, *Moloi* (witch/wizard) 1618 and *ditaola* (divination bones) at 521. Second, *Jeso* (Jesus), *Baebele* (Bible) and *Bakeresete* (Christians) do not make it into the top 5,000 prose text. A look at the TSB terms also reveals that *badimo* does not make it into the top

5,000 words of the combined text while *setlhabelo* and *matwetwe* do not make it to the prose text. Therefore in the 12 words that we have inspected in the two tables, 5 of the words do not make it to the top 5,000 prose text and only one does not make it to the top 5,000 combined list. These results are significant in that they reveal that almost half of the inspected words do not make it into the top 5,000 words of prose text.

Lexicographically, the implications are serious. Missing words in a dictionary such as the ones inspected above leads to gaps in the lexical representation of a language in a dictionary.

We now look at the grammar text and inspect some basic grammatical terms and measure the performance of both lists. Grammar texts are studied since they are central to students' Setswana grammar classes which are compulsory at both junior and senior secondary schools. Basic grammatical terms would therefore be expected in school dictionaries, even short ones. Below we present the results of six grammatical terms.

Table 83: Grammar terms and their position on the two lists

Setswana	English	Prose text	Combined text
Tumanosi/ditumanosi	Vowel(s)	-	35911389
Lediri	Verb	884	898
Tumammogo/ditumammogo	Consonant(s)	-	43502247
Letlhaodi	Adjective	4,544	3089
Letlhalosi	Adverb	3,409	2125
Leemedi	Pronoun	-	4569

The results above show that half of the words do not show up in the most frequent 5,000 words of prose text. These are *tumanosi* (vowel) and its plural *ditumanosi* (vowels), *tumammogo* (consonant) and its plural *ditumammogo* (consonants) and *leemedi* (pronoun). The results reveal a lack of some of the basic grammatical labels in the prose text. On the other hand, all the words inspected appear in the combined list. These results are consistent with the previous results where about half of the words do not appear in the restricted list but do appear in the combined list. The results show that while prose texts deal with a variety of subjects they have limitations

when specialised areas like linguistics are studied.

We also look at the business terms and how they perform in both the prose and combined list. The results follow in Table 84:

Table 84: Business terms and their rank on the two lists

Setswana	English	Prose text	Combined text
Bagwebi	Business people	-	787
Kgwebo/dikgwebo	Business/businesses	1151/3978	585/1094
Mmaraka	(market	-	1133
Madirelo	Factories	3935	933
Kompone	Company	-	629
Itsholelo	Economy	-	679

Only two of the six business terms make it into the top 5,000 prose words. The two words that do make it into the top 5,000 are comparatively ranked lower in the list. The business terms results are consistent with the results which have been seen so far with grammar and religious terms from Christianity and traditional Setswana beliefs.

For our final measurement we look at taboo words; insults or vulgarities which rarely make it into school textbooks, local newspapers and dictionaries. Landau (1989: 187) laments that, "[n]o aspect of usage has been more neglected by linguists and lexicographers than that of insults." Their lack of inclusion in such texts is barely surprising since insults are not just taboo, but by their nature they constitute what Butler (1997: 2) calls "injurious speech" or signs used with the intention to shock, bring offence and psychological harm to the targeted individual or group. Insults are therefore instances of linguistic violence; reflections of how individuals verbally inflict injury on each other (cf. McEnery and Xiao, 2003). They take different forms. Some refer to private parts while others are rude words which refer to embarrassing actions particularly when mentioned in public (also see Lynch, 2004: 640[28]). These actions may include references to relieving oneself or they may take the form of coarse words referring to sexual activity or farting. They may also be group insults

[28] Lynch lists insults from Johnson's dictionary which include amongst others: *airling, asshead, backbiter, backfriend, barbarian, bedpresser, bellygod, bitch, blockhead, blowze, blunderhead, booby, barachio, bufflehead, bumpkin, bungler, simpleton, noodle* and *smellfeast*

which refer to and label particular ethnic groups or a particular sex.

There are a variety of reasons why such language is relevant to academic study and more so to have a place in dictionaries. Such reasons include amongst others:

i. Adult learners of a new language, or those who have moved into a new society with a totally different language, may take a keen interest in knowing rude words as a defence mechanism – so that they may be aware when insults are hurled at them.
ii. Insults are taboo; therefore an understanding of insults will contribute to an understanding of a society's taboos – an understanding of what is socially acceptable or profane.
iii. As stated previously, profanities may be perceived as cases of linguistic violence – inflictions of injury on the other. In this way a study of insults may be seen as a study of social violence.
iv. Users search for insults in dictionaries. De Schryver and Joffe in their study that makes a determination of how electronic dictionaries are used. They have found out that,

> [i]n the top 100 searches there are a further 6 foreign words lows (4 Setswana and 2 English), and of the remaining 31 words no less than 17 either have to do with the sexual sphere or are extremely offensive: *marêtê* 'testicles', *masepa* '(*off.*) shit', *mogwêtê* '(*off.*) anus', *mpopo* '(*off.*) private part (vagina; penis)', *nnyô* 'vagina', *nnywana* '(*off.*) cunt', *ntoto* 'penis', *nyôba* '(vulgar) fuck', *sefêbê* 'prostitute; (off.) bitch', *thôbalanô* 'sex', etc. This latter phenomenon might very well be the case for all (Internet) dictionaries (De Schryver and Joffe, 2004: 190).

They also observe that,

> An analogous study of the top 100 English searches reveals a similar pattern, with 18 of the top 100 searches also in the BNC top 100 (Leech et al. 2001) and 62 in the BNC top 1 000. A single item in the

top 100 searches is misspelled, while 6 of the remaining 37 searches again belong to the same sexual/offensive sphere: *bitch*, *fuck, penis, sex, shit* and *vagina* (De Schryver and Joffe, 2004: 190).

De Schryver and Joffe's findings give support to the study of insults as an interesting academic area of investigation.

However the point of this section is not an attempt at a study of profanities but rather to use the absence or presence of insults as an illustration of the strength or weakness of a corpus text type coverage. The point is that if corpora are based on texts which have been edited by publishers and newspaper editors who may be following prescriptive rules about a language, then a corpus itself may be only offering a partial reflection of the state of a language. That is why in the design and compilation of the Setswana language corpus we have incorporated chat-site material. While the material has greater levels of English words, it does provide valuable Setswana language style that is rarely seen in published texts but is characteristic of youthful dialogues. Vulgarities and words that refer to private parts while frequently avoided by publishers do occur in chat-site material.

We therefore look at the different vulgarities and present the results in Table 85:

Table 85: Vulgarities and their position on the two lists

Setswana	English	Prose text	Combined text
Marete	balls, scrotum	-	2269
Polo	dick, penis	-	2087
Masepa	shit	-	1725
Nnyo	vagina	-	982
Phona	vagina	-	-
Sebono	asshole, anus	-	2115

Prose text does not provide any evidence of any of the common vulgarities that we have isolated. This is barely surprising since most of the Setswana prose is primary school, secondary school and university educational material which sanctions vulgarities. Combined texts on the other hand show very high presence of vulgarities.

The different experiments above in which we compare the 5,000 word prose list and

5,000 word combined list aimed at comparing the performance of an opportunistic corpus against a broad coverage corpus. The results of all the experiments point to the inadequacy of an opportunistic corpus (in this case, a single text type corpus) as a reliable source of dictionary material. They reveal that the simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus.

7.7 Chapter conclusion

In this chapter we have explored a variety of experiments to determine if a corpus comprising a variety of text types was any different from one with a single text type on the basis of the types it contributed at comparable intervals. We began by first segmenting the Setswana corpus into 10,000 token chunks. For every text type, the types' measurements were taken at 10,000 token intervals. The 10,000 token-chunks were randomised for every measurement taken and the experiment iterated five times. The type measurements were taken at every 10,000 token interval up to 500,000 tokens. An average was computed so that comparisons between text types using a single mark that summarises the results at every 10,000 tokens interval could be made.

The experiments revealed Poetry text as having the largest number of types at most of the 10,000 tokens intervals followed by PRONEWHANCAL and SCIPOLBUSREL. We have also found out that the combination of text from a variety of text types compiled into POEGRACHAPLA, SCIPOLBUSREL and PRONEWHANCAL resulted with higher types when compared to the distinct text types from which their parts were compiled.

The experiments also revealed that Politics text had the lowest types overall, followed by Call-in and Chat-site texts. This suggests that these three use a limited vocabulary when compared with other text types. We argued that while certain text types contribute the lowest number of types, such a smaller number of types should not be perceived as implying less importance or less significance in corpus compilation, since even the text types with the lowest number of types do contribute unique words to other text types.

The performance of the most frequent 100 words from different text types was measured against the most frequent 100 words of the whole Setswana corpus. It was found out that it was not enough to just have a corpus with a variety of text types to generate large numbers of types. It was also crucial that the individual text types that comprise a corpus should individually have large numbers of word types.

Simple consistency analysis (SCA) which calculates dispersion or word-spread in corpora was also explored. The SCA results were compared to the calculation of raw frequencies in the calculation of the most frequent words in the corpus. SCA has been able to determine whether a widely spread use of a word is because it occurs in many text samples or whether it is frequent because of high usage in only a few texts. The SCA calculation computes words which recur consistently in texts and orders them on the basis of their spread across documents.

The most frequent words in different text types were compared. Raw frequencies were chosen in the comparison of the most frequent 100 words so that the results could be comparable to those of other wordlists of other corpora such as the BNC.

We also compared the most frequent 100 words of the written component of the BNC and the most frequent 100 words of the spoken BNC component against the most frequent 100 words of the whole BNC. The results of this experiment were compared with the Setswana corpus experiment. Seventy one of the Setswana spoken subcorpus' most frequent 100 words were found amongst the most frequent 100 Setswana words. Eighty one of the most frequent 100 written words were found amongst the most frequent 100 words of the complete Setswana corpus. The BNC, on the other hand had 97 of the top 100 words of the written component of the corpus in the most frequent 100 words of the whole corpus and 72 words of the top 100 words of the spoken component. The Setswana corpus therefore compared well with the BNC corpus in this experiment.

To further test whether text type diversity was crucial to the kind of words which are selected for inclusion in a dictionary, two 5,000-word list chunks were compared. The first chunk simulated an opportunistic corpus with its text type limitations since it was

derived exclusively from the Prose text. Prose text was chosen since many readily available text materials in African languages are of a prose type which would comprise novels. The most frequent 5,000 words were therefore derived from the prose text and compared with 5,000 words compiled from a variety of text types. Both lists were tested for a variety of grammatical, religious and business terms and for certain Setswana vulgarities. The results showed that the simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus. It was also found out that some of the most frequent words in Setswana were found in both corpora. The results provide evidence for broad text type coverage in corpora compilation as a reliable source of broad lexical coverage for dictionary compilation.

What the different experiments have shown is that there are considerable differences between the different wordlists extracted from the diverse text types. The experiments testify to the limitation of a single text types as a source of dictionary evidence. They have shown that to get a variety of words of a language, a corpus with diverse text types is preferable.