Chapter 6

Measuring text type diversity

6.1 Introduction

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

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

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

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

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

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

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

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

We begin by looking at keyword analysis.

6.2 Keyword analysis

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

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

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

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

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

and these are cross-tabulated.

One way of explaining this process is to say:

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

2. Generate frequency lists from the two subcorpora.

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

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

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

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

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

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

The statistical tests include:

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

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

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

Table 31: A contingency table

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

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

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

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

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

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

This equates to calculating log-likelihood G2 as follows:

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

Kilgarriff notes that,

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

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

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

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

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

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

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

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

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

Below we extensively use log likelihood measurement in keyword analysis to isolate words which characterise a genre or text type. Our aim is two fold: at one level we wish to measure whether a corpus compiled from various text types and genres is more attractive for lexicography if it could be found to generate unique words particular to each genre, which collectively capture the linguistic diversity present in every day language. A corpus that is designed in such a way as to capture the linguistic diversity of a language would therefore be preferred over a corpus compiled from a single or a limited variety of genres. At another level, we hope that keyword analysis lists will by their distinctiveness communicate a related argument: that since the lists are dissimilar the course of lexicography cannot be served best by depending on a single text type for dictionary compilation, since a single text type will lead to the generation of a restricted lexicon. We use keyword analysis to extract genre specific lists, which Vintar (1999: 64) has argued that they "… can prove useful when studying the lexical specificity of a text or its terminological scope," an area we are currently investigating.

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

6.2.1 Keyword analysis of written components of the Setswana corpus

Table 32: Science and technology keywords

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

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

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

The Science and technology text type has comparatively higher levels of English words compared to the other lists that we will inspect later. Amongst these are *workshop* (64), *radio* (78), *rabies* (81), *technology* (83) and *telecommunications* (92). This is in part because some documents written in Setswana use English terms where the Setswana language does not have terms for certain science and technological concepts. In other instances the English words have been adopted into the Setswana language and spelt using Setswana orthography. Instances such as *inthanete* (20) (internet), *nyutlelia* (nuclear) (39), *yunibesithi* (university) (59), and *dikhomputara* (62) (computers) are examples of Setswana words borrowed from English.

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

Table 33: Politics text keywords

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

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

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

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

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

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

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

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

Table 35: Poetry text keywords

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

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

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

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

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

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

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

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

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

Table 36: Plays text keywords

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

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

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

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

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

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

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

Table 37: Plays text keywords with names treated as metatext

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

77.	kampo	618.20
78.	sengwe	604.83
79.	be	602.51
80.	ee	590.91
81.	tshega	586.26
82.	fano	565.70
83.	sentle	563.01
84.	tswa	557.95
85.	tlile	548.69
86.	twe	548.02
87.	ao	545.37
88.	ene	540.97
89.	utlwile	534.76
90.	botlhoko	528.49
91.	mosetsana	522.90
92.	tlhogo	522.21
93.	jang	519.62
94.	bosigo	514.07
95.	setse	501.38
96.	ise	491.63
97.	ijoo	486.10
98.	ena	485.28
99.	dikgomo	483.85
100	ntlong	475.17

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

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

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

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

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

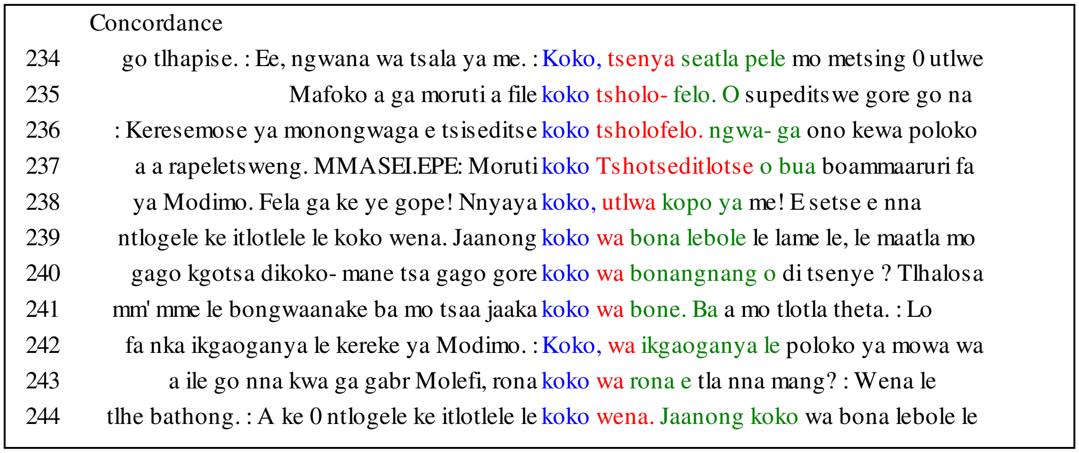
Koko (74) in Setswana is an ambiguous word since it could mean "chicken" or it could be "a verbal knock" at the door or a short form for "*nkoko*" (grandmother). A concordance analysis of the whole corpus revealed 249 concordance lines. Two instances of these are with *koko* as a knock at the door. This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

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

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

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

The overwhelming uses of "*koko*" are as a shortened version of *nkoko* (granny). We offer a few concordance lines to illustrate such usage:



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

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

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

Table 38: Grammar texts keywords

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

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

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

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

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

Table 39: Arts & culture text keywords

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

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

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

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

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

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

Table 40: Chat-site text keywords

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

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

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

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

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

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

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

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

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

Table 41: News text keywords

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

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

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

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

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

[22] Nicknames for the Botswana Democratic Party

[23] Botswana National Front

[24] Botswana Telecommunications Authority

[25] Information Communication Technology

Table 42: Religious text keywords

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

100.lotso

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

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

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

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

6.2.2 Keyword analysis of spoken components of the Setswana corpus

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

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

Table 43: Call-in text keywords

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

8	hello	1742.438477
9	jaanong	1212.748047
10	nnyaa	1009.634583
11	tankie	720.6206055
12	raya	703.6761475
13	mma	637.4937744
14	ntse	590.5014648
15	ehe	589.7432861
16	lapeng	564.6315918
17	okay	530.8041382
18	ngwana	524.657959
19	cellphone	523.3270874
20	ipolaya	512.5631714
21	rre	507.0570068
22	kana	499.7451172
23	fa	497.8006287
24	bo	496.7189941
25	lebogile	470.307373
26	nna	461.8958435
27	radio	452.9598694
28	tla	452.5892639
29	bye	428.1827087
30	le	415.0164795
31	bua	403.5699768
32	teng	403.0974426
33	wena	392.680542
34	bona	391.37854
35	bana	389.4447021
36	jang	387.3610535
37	tota	367.9024658
38	mme	366.5827637
39	fela	354.6725464
40	batsadi	344.4155884
41	ntate	340.0013428
42	leng	335.1376953
43	rona	333.7392578
44	hela	325.0430603
45	gone	310.8773499
46	eng	305.8117371
47	nkgonne	304.1421204
48	itse	295.0811462
49	mo	278.6512756
50	utlwa	270.1494751
51	go	269.0933533
52	ga	262.1260986
53	batla	249.0783386
54	na	248.7524261
55	mogaetsho	248.1150818
56	jaana	245.1821289
57	pule	243.3664246
58	gongwe	238.5174866
59	ba	229.0778656
60	rraagwe	217.4441681
61	leboga	212.4922943
62	ngwanaka	210.8857269
63	matlhophelong	206.3597717
64	kae	196.0868378
65	moso	195.6818695
66	ka	194.9997406
67	ra	190.0220947
68	ha	188.3122253
69	chris	186.6708221
70	gago	183.3898163
71	golo	182.3194733
72	reng	180.725174
73	phutha	173.7875977
74	se	167.048996
75	lo	166.1824951
76	tsaya	164.6233063
77	moremogolo	163.4777679
78	ye	158.3822937
79	ditshaba	156.4868469
80	dilo	152.937088
81	ehee	148.8966675
82	tsamaya	148.3753815
83	semang	143.8077393
84	kgang	142.1442719
85	gaetsalwe	142.0330048
86	gompieno	140.5166931
87	sengwe	140.3969421
88	utlwile	136.7749939
89	rraetsho	135.7851563
90	nale	132.7247314
91	rekela	132.4711914
92	motsadi	127.2045517
93	thupa	124.4304581
94	kgona	121.6762238
95	dithato	121.6504059
96	tle	121.1425247
97	bue	120.5993958
98	setlhogo	120.2457581
99	mozeregwa	119.606102
100	campaign	117.7104187

The list gives evidence of words common in dialogue by the use of such words as *ee* (1) (yes), *nnyaa* (10) (no) *hello* (8), *tankie* (11), or *ke lebogile* (2/25) (thank you) and *bye* (29).

Pronouns in direct communications rank amongst the most frequent. These are *nna* (26) (me), *wena* (33) (you), and *rona* (43) (us).

A variety of words which mark interrogatives appear in the list indicating that the presenter asks a series of questions to the callers. These include words *jang* (36) (how) *leng* (42) (when), *eng* (46) (what), *kae* (64) (where/how much).

Words that reflect respectful dialogues between individuals *nkgonne* (47) (elder brother/sister), *mogaetsho* (55) (colleague), *batsadi* (40) (parents) *rraetsho* (89) (sir), *motsadi* (92) (parent), appear amongst the top 100.

Other words hint at the topics discussed in the dialogues. Some of these words are *cellphone* (19), *ipolaya* (20) (commit suicide), *matlhophelong* (63) (voting stations), *rekela* (91) (buy for) and *thupa* (93) (lash/stick).

Included in the spoken subcorpus are the face-to-face dialogues. These are recordings of family interactions. The text from these is small. Consequently only 71 keywords have been extracted instead of 100.

Table 44: Face to face dialogue keywords

N	Keyword	Keyness
1	re	490.53
2	ee	338.58
3	nnyaa	208.44
4	ke	203.4
5	ko	142.62
6	ba	137.33
7	lo	129.81
8	kgosing	117.72
9	raya	106.36
10	ehe	99.175
11	plaka	91.315
12	fela	80.89
13	mmathapelo	78.269
14	tswaletswe	71.928
15	dikgomo	70.978
16	mm	65.223
17	tlhotse	64.623
18	hela	64.497
19	tswetswe	58.268
20	malutu	57.128
21	kana	55.318
22	ga	54.396
23	win	52.177
24	ijoo	51.33
25	gana	51.03
26	bogadi	48.315
27	lona	48.075
28	talela	47.176
29	apeetse	47.176
30	fologa	44.879
31	welang	43.33
32	tshekong	42.774
33	montshonyana	42.626
34	tswale	41.099
35	kere	39.627
36	mpoleleleng	39.132
37	leteitsi	39.132

38	theogela	39.014
39	molaodi	38.759
40	jaana	36.09
41	tsee	35.727
42	sekgoa	34.125
43	gatwe	32.629
44	ha	31.573
45	maabane	31.263
46	dikologa	30.98
47	golo	30.411
48	itshwere	29.917
49	mmm	29.583
50	bo	29.582
51	eng	29.426
52	mpolelele	28.562
53	twe	28.242
54	fa	28.138
55	jaanong	28.088
56	tv	27.887
57	tleng	27.887
58	kwano	27.771
59	tilwe	27.693
60	gore	27.684
61	nyalwa	27.392
62	gogwa	26.926
63	tweng	26.891
64	batla	26.88
65	nna	26.86
66	phakela	25.121
67	nyetse	25.121
68	bogosi	25.118
69	phoso	24.326
70	yo	24.23
71	letse	24.115

As in many dialogue instances, the face to face dialogue text has many functional terms such as pronouns. These are: *re* (1) (we), *ke* (4) (I), *ba* (6) (they), *lo* (7) (you), *lona* (27) (you), *bo* (50) (it/those), *fa* (54) (here/give), *kwano* (58) (here), *nna* (65) (me/sit) and *yo* (70) (this one). Pronouns signal individuals' close interaction with their immediate environment as they point and make reference to where they are.

Other marks of personal interaction are expressed in reactions to what is being said. Such terms include words such as *ee* (2) (yes), *nnyaa* (3) (no), *ehe* (10) (wow/I see), *mm* (16), *mmm* (49) and *ijoo* (24) (interjective of surprise or shock),

Other words that signal interlocutors who are engaging each other are: *ko* (5) (at), *fela* (12) (only), *tlhotse* (17) (spent the day), *hela* (18) (only), *kana* (21) (or), *ga* (22) (of), *jaana* (40) (this way), *eng* (51) (what), *tilwe* (59) (said), *gore* (60) (that), *batla* (64) (want).

As part of the fieldwork, classroom interactions were recorded at junior secondary schools. Since Setswana is used in the teaching of Setswana grammar and literature, only Setswana text from such classes has been recorded, transcribed and added to the corpus. The top 100 keywords from the text are in Table 45.

Table 45: Educational spoken text keywords

N	Keyword	keyness
1	ee	1658.3
2	ke	1384.3
3	letlhalosi	1146.4
4	lebopi	854.85
5	re	823.11
6	sekai	821.37
7	eng	769.09
8	utlwana	726.59
9	felo	679.43
10	kere	590.38
11	ra	496.72
12	dikai	492.56
13	ko	484.48
14	gore	473.02
15	ehe	461.73
16	thito	456.45
17	ga	451.52
18	le	441.43
19	mphang	430.18
20	letlhaodi	427.34
21	raya	422.33
22	hee	402.19
23	popo	376.14
24	ngotlo	333.99
25	kae	323.02
26	mpha	314.13
27	rra	284.99
28	modumo	266.09
29	go	259.6
30	mma	252.9
31	tengwafatso	252.5
32	akere	234.56
33	lerui	233.08
34	letshwaogoka	233.08
35	seyantlo	233.08
36	leemedi	229
37	mphe	223.77
38	o	220.12
39	jaanong	217.85
40	ile	217.2
41	kana	215.64
42	mabopi	213.65
43	masimo	207.92
44	tsamaela	206.38
45	wena	205.7
46	mogatlana	205.44
47	lenyalong	197.21
48	malome	187.78
49	nnyaa	180.37
50	gokelela	176.46
51	fe	175.36
52	ditlhaodi	174.8
53	lesoboki	174.8
54	na	171.63
55	nnya	171.22
56	dirisitse	167.67
57	araba	157.38
58	rinifatso	155.38
59	pirwana	155.38
60	mefuta	149.71
61	mosadi	147.41
62	lone	146.65
63	tlholego	145.3
64	lenyalo	144.02
65	rile	143.03
66	bakang	142.17
67	mothofaditsweng	135.96
68	leamanyi	135.96
69	poufatso	135.96
70	la	132.91
71	leina	131.77
72	bua	129.55
73	itse	123.81
74	sengwe	123.58
75	kwala	122.63
76	utlwe	122.44
77	lengwe	120.26
78	reng	119.77
79	supa	116.
80	lefelo	115.92
81	kedibonye	114.13
82	tsholetsa	113.77
83	tlhaloso	111.59
84	fa	111.32
85	waitse	110.85
86	gago	110.02
87	fela	109.63
88	tilodi	109.33
89	yoo	108.63
90	potso	108.53
91	bona	108.19
92	jang	106.95
93	tsamaetse	105.31
94	gona	104.79
95	lebaka	104.51
96	akanya	103.87
97	dirise	103.82
98	efe	102.48
99	letsogo	101.59

100	tsweleleng	101.57

Since most Setswana classes deal with Setswana linguistics this is reflected in the grammatical labels that are captured in the Table 45 list. These include *letlhalosi* (3) (adverb), *lebopi* (4) (morpheme), *thito* (16) (stem), *letlhaodi* (20) (adjective), *ngotlo* (24) (diminution), *modumo* (28) (sound), *tengwafatso* (31) (palatalisation), *leemedi* (36) (pronoun), *mogatlana* (46) (suffix), *lesoboki* (quantitative), *rinifatso* (lateralisation), *leamanyi* (relative), *mothofatso* (personification), and *poufatso* (labialisation). Setswana classes in general teach Setswana grammar, culture and literature. That is why there are many grammatical terms. Other terms give signal to the giving of instructions found in instruction classes. These include amongst others *sekai/dikai* (6/12) (example/examples), *mphe* (37) (give me), *gokelela* (50) (link/connect), *araba* (57) (answer), *kwala* (75) (write), *supa* (79) (show), *tlhaloso* (83) (explanation), *potso* (90) (question), *akanya* (96) (think), and *efe* (98) (which one?) Setswana cultural terms include *utlwana* (8) (be at peace with), *masimo* (43) (farms), *lenyalong* (47) (concerning a wedding), *malome* (48) (uncle), *pirwana* (59) (black colour of a female sheep), *mosadi* (61) (woman), *bakang* (66) (cause/praise), and *tilodi* (88) (black and white animal colour).

The largest part of the spoken subcorpus is made of Hansard text. The text was scanned from Hansard publications from the Botswana parliament. The most frequent 100 keywords are presented in Table 46.

Table 46: Hansard spoken text keywords

N	Keyword	Keyness
1	gore	20560
2	re	17086
3	ke	10741
4	i	9011.6
5	the	8464.9
6	mr	8217.9
7	of	5400.2
8	to	5328.6
9	leng	4743.7
10	jaanong	4665.1
11	ko	4624.6
12	rraetsho	4421.4
13	hansard	4149.3
14	page	4096.8
15	honourable	3876.3
16	that	3777.9
17	speaker	3648.4
18	is	3512.9
19	and	3342.6
20	bua	3210.8
21	ba	3070.2
22	motsamaisa	2953.9
23	debate	2757.2
24	member	2733.3
25	you	2670.8
26	resumed	2634.3
27	we	2506.8

28	palamente	2392.7
29	motion	2248.4
30	dipuisanyo	2238.5
31	rona	2101.4
32	mme	2091.4
33	bo	2087.6
34	bill	2039.1
35	in	1922.5
36	not	1745.8
37	have	1682.7
38	kana	1673.1
39	are	1672.1
40	this	1670.6
41	it	1664.3
42	ra	1622.6
43	fa	1439.8
44	di	1415
45	march	1411.8
46	teng	1411.1
47	jaana	1410.2
48	tona	1401.4
49	nnyaa	1395.1
50	appropriation	1350.1
51	minister	1337.7
52	yone	1313.4
53	second	1266.1
54	be	1247
55	gone	1209.2
56	reading	1201.5
57	go	1181.5
58	for	1135.4
59	leboga	1109.2
60	motlotlegi	1088.9
61	order	1088
62	se	1085.7
63	on	1084.2
64	draft	1071.7
65	ka	1071
66	point	1054.8
67	gona	1052.3
68	tse	1050.3
69	fela	1029.4
70	ntseng	1018.3
71	development	1015.2
72	gongwe	1013.3
73	buang	955.65
74	privatisation	948.36
75	ga	946.16
76	policy	924.36
77	they	909.84
78	batho	898.65
79	goromente	888.41
80	tsone	874.72
81	rra	861.48
82	ntse	861.1
83	tie	857.75
84	one	850.56
85	head	848.72
86	tuesday	833.77
87	dilo	828.97
88	wednesday	818.69
89	speech	814.07
90	thursday	807.58
91	eleng	798.9
92	clarification	798.86
93	batswana	791.79
94	as	787.79
95	what	781.11
96	there	767.99
97	itse	755.23
98	but	752.49
99	monday	752.22
100	ministry	746.5

The Hansard as an official report of parliamentary speeches has a preponderance of formal parliamentary terminology such as *Mr*. (7), *rraetsho* (13) (Sir), *honourable* (16) or *Motlotlegi* (65), *speaker* (18) or *Motsamaisa* (23) *Dipuisanyo* (31) in Setswana, *member* (25), *palamente* (29) (parliament), *motion* (30), *bill* (35), *Minister* (52), point (71) of (8) order (66), reading (60), draft (69), and *goromente* (86) government. The subcorpus does display common terms in speech such as *gore* (1) (so that), *re* (2) (we), *ke* (2) (I), *the* (5), *of* (7) to (8), *leng* (9) (when), *jaanong* (10) (now), *ko* (12) (at) *that* (16) is (18), it (41) *yone* (52) or *gone* (55) or *gona* (67), in (35), and *not* (36). Other terms like *appropriation* (50), *development* (71), *privatisation* (74), and *clarification* (92) indicate the high register which often characterise parliamentary debates. What may be observed is the high occurrence of English terms even in what is Setswana text. There are at least two explanations for this. First, English is an official language in Botswana and educated speakers tend to code switch freely, particularly in official contexts such as parliamentary debates. Second, parliamentary debates make use of specialised terminology which Setswana language has not been developed to handle adequately. Some of the instances of English terminology usages are *bill* (34), *in* (35), *not* (36), *have* (37), *March* (45), *appropriation* (50), *second* (51), *be* (54), *reading* (56), *for* (58), Monday (99), and *ministry* (100).

Spoken subcorpus also comprises a television interview from the Botswana television (Btv) program, *The Eye*. The recorded program was about water conservation in Gaborone in light of the 2004 drought which nearly dried the Gaborone dam which supplies the city of Gaborone with water.

Table 47: Interviews spoken text keywords

N	Keyword	Keyness
1	metsi	2339.9
2	gore	2202.7
3	re	2064.4
4	ke	1664.1
5	ee	1252
6	leng	1249.5
7	eh	1047.5
8	water	955.97
9	ehe	885.45
10	mma	869.78
11	le	677.77
12	ko	386.17
13	utilities	338.61
14	na	302.3
15	eng	292.63
16	bo	272.35
17	rona	271.35
18	dam	261.11
19	letsibogo	258.01
20	mang	223.55
21	be	216.35
22	gone	215.13
23	demand	203.92
24	fem	203.92
25	one	198.55
26	ga	197.75
27	kana	188.73
28	go	186.67
29	so	182.44
30	dintshu	169.93
31	affairs	169.52
32	mme	169.37
33	raya	167.7
34	ra	162.9
35	pipe	161.92
36	fela	161.34
37	nnyaa	156.86
38	dikoko	156.32
39	bua	154.24
40	itse	151.05
41	jaana	150.02
42	lehuma	148.08
43	supply	145.13
44	technology	140.52
45	ka	140.13
46	mo	138.08
47	map	136.66
48	mananeo	133.51
49	fifty	132.19
50	the	130.54
51	jaanong	125.67

52	tweng	123.17
53	matamo	122.19
54	di	122.12
55	waste	119.46
56	litres	118.95
57	letamo	116.44
58	nosa	114.43
59	tlhatswa	112.87
60	pit	106.95
61	nnya	106.69
62	gago	106.55
63	dirisa	106.36
64	femp	101.96
65	menoto	101.96
66	latrine	101.96
67	gaborone	101.49
68	jang	99.687
69	batswana	98.081
70	rre	95.613
71	carrier	94.935
72	nna	94.054
73	dirise	91.167
74	pompa	90.531
75	ceda	89.293
76	ao	89.109
77	tamong	87.031
78	biditswe	86.29
79	tse	86.233
80	lone	85.584
81	maybe	84.964
82	mathata	83.488
83	teng	81.18
84	lona	80.876
85	ba	78.587
86	dilo	76.233
87	tla	76.096
88	tlase	75.679
89	sentle	75.489
90	gape	74.288
91	eight	74.208
92	conservation	74.208
93	tamo	74.208
94	mmitsa	72.787
95	two	72.635
96	kgona	71.154
97	gogwe	71.005
98	metseng	70.544
99	four	70.432
100	corruption	69.994

The subject dealt with in the interviews is clearly revealed by the terms that are key. These include *metsi* (18) (water), (water) *utilities* (13) (the water provider in cities), *dam, letamo, tamo* (18, 57, 93) (dam), *Letsibogo* (19) (the name of a dam), *demand* (23), (water) *affairs* (31) (the water provider in villages), *pipe* (35), *supply* (43), *technology* (44), *waste* (55), *litres* (56), *nosa* (58) (serve water), *tlhatswa* (59) (purify), *mathata* (82) (problems), *conservation* (92).

What stands out as well in this list is a high level of English usage since the subject is technical. These include carrier *water* (8), *utilities* (13), *demand* (23), (71), *maybe* (81), *eight* (91), *conservation* (92), *four* (99), and *corruption* (100).

The Open radio programming subcorpus includes a variety of different radio programs. Amongst these are *Matimela* (a program about lost and found cattle), *Tatediso ya dikgang* (a news program that follows evening news featuring reports from reporters from around the country), *Borukutlhi* (an anti-crime program), and *Molemi-ithute* (an educational program for farmers). The subcorpus is therefore diverse in its coverage since *Tatediso ya dikgang* as a news program covers a variety of subject matters. Because of the two programs for farmers there are many agricultural terms in Table 48 list.

Table 48: Open radio programming keywords

N	Keyword	Keyness
1	ba	857.19
2	ko	655.83
3	tshipi	440.85
4	tshwailwe	401.26
5	le	318.42
6	boitaolo	258.23
7	re	228.15
8	tlhomagane	206.36
9	di	194.94
10	ya	194.87
11	tlhaka	193.51
12	go	169.04
13	pelesa	168.32
14	serope	162.18
15	mojeng	148.91
16	molemeng	145.57
17	sekolong	145.28
18	baithuti	134.17
19	ke	126.12
20	gore	123.15
21	mo	111.68
22	ka	110.39
23	superintendent	105.07
24	bana	104.62
25	tse	103.41
26	la	98.832
27	kgomo	98.613
28	mathateng	98.369
29	khunwana	88.834
30	mapodisi	85.007
31	sekolo	82.43
32	kgaolong	80.714
33	moroba	77.782
34	batsadi	76.667
35	khamphane	76.652
36	lephaga	76.415
37	maphaga	76.415
38	wa	76.299
39	bao	75.991
40	tlase	73.234
41	nako	72.476
42	godimo	72.133
43	eo	69.95
44	bone	69.666
45	na	68.534
46	batshabi	66.353
47	kwena	65.801
48	bakgweetsi	62.958
49	ngwe	62.165
50	dikgang	62.085
51	Hiv	60.917
52	Te	59.219
53	foods	57.667
54	tsa	57.532
55	selebi	56.715
56	matshwao	56.715
57	aids	52.997
58	bag	52.713
59	isa	52.512
60	tsela	52.492
61	bese	51.04
62	bo	51.04
63	bdp	50.496
64	mapodise	50.307
65	party	47.854
66	kgongwana	47.758
67	tsenya	47.532
68	mme	47.233
69	phati	46.948
70	fitileng	46.59
71	diphologolo	46.304
72	maswabi	46.235
73	khunou	45.386
74	babelaelwa	45.386
75	kopa	44.905
76	kgabaganya	44.854
77	dipalo	44.737
78	matimela	43.918
79	leng	43.098
80	dijo	42.797
81	dira	42.613
82	mokgweetsi	42.456
83	tatediso	42.237
84	rile	41.965
85	neng	41.464
86	lwetse	41.343
87	rre	40.758
88	fa	39.813
89	khampane	39.595

90	theme	39.417
91	mosong	39.327
92	pharakano	38.75
93	tshologa	38.272
94	mokgaoganyi	38.207
95	cosatu	38.207
96	tla	37.933
97	mmuso	37.786
98	dikotsi	37.718
99	jaana	37.281
100	constable	37.272

Because of the subjects handled by the programs, there are terms related to branding of cows such as, *tshipi* (3) (metal used for branding cows), *tshwailwe* (branded), *tlhaka* (11) (branded letter), *serope* (14) (thigh, where cows are branded), *mojeng* (15) and *molemeng* (16) (right and left side; sides on which cows are branded), *lephaga/maphaga* (36/37) (a type of animal ear mark), and *matshwaô* (56) (marks). Other terms refer to the kind and or size of the animal. These include, *pelesa* (13) (heifer), *kgomo* (27) (cow), *moroba* (33) (mid-sized cow), and *kgongwana* (66) (calf). Other terms refer to the colour of the cows. These are *khunwana* (29) (reddish brown on female cows) and *khunou* (73) (reddish brown on male cows).

Other terms point to crime prevention and police work. These include *boitaolo* (6) (rebelliousness), superintendent (23), *mapodisi/mapodise* (30/64) (police officers), *babelaelwa* (74) (suspects), *mokgweetsi/bakgweetsi* (82/48) (driver(s)), *dipalo* (77) (statistics/numbers), *pharakano* (92) (traffic), *dikotsi* (98) (accidents) and constable (100).

Other words are educational. These include *sekolong/sekolo* (17/31) (of school/school), *baithuti* (18) (learners). There are a variety of terms which probably come from different news items. These include *khampane* (35) (company), *nako* (41) (time), *godimo* (42) (above/on top), *batshabi* (46) (refugees), *kwena* (47) (crocodile), Aids (57), bag (58), *tsela* (60) (way/road), *bese* (61) (bus), BDP[26] (63) (the Botswana ruling party), *kgabaganya* (76) (cross), *mosong* (91) (morning), *mmuso* (97) (government).

There is also a considerable use of pronouns such as *ba* (1) (of), *ko* (2) (at), *re* (7) (we), *mo* (21) (in), *tse* (25) (these), *bao* (39) (those), *eo* (43) (that one), *bone* (44) them, *fa* (88) (here). These are common in spoken language.

[26] The Botswana Democratic Party

The religious spoken text is exclusively from the Christian faith. The other's faiths have a very small following nationally (Christian 71.6%, Badimo 6%, other 1.4%, unspecified 0.4%, and none 20.6% (The Republic of Botswana: Central Statistics Office, 2001 census)). The data comprises sermons from churches and funerals and from the radio program, *Sidilega* (be well). The keywords from this data follow in Table 49.

Table 49: Religious spoken text keywords

N	Keyword	Keyness
1	Re	1257.1
2	ke	1125.4
3	modimo	661.24
4	mme	412.19
5	bagaetsho	392.82
6	gago	378.96
7	amen	296.5
8	leboga	293.06
9	le	293.02
10	lefoko	287.22
11	mma	284.13
12	lo	270.29
13	eh	236.18
14	yo	232.77
15	morena	201.33
16	rra	195.2
17	jaanong	185.97
18	gore	181.87
19	tle	177.6
20	tlaa	159.44
21	ga	147.87
22	ka	140.81
23	mo	139.63
24	batsadi	137.04
25	galalelang	136.11
26	bua	130.29
27	bagaetshong	127.95
28	kwano	127.53
29	nne	124.24
30	baruti	123.74
31	rona	114.67
32	be	107.29
33	kagiso	105.6
34	ngwana	105.02
35	fano	104.1
36	bo	103.27
37	jaana	101.1
38	fa	100.6
39	keresete	99.279
40	sefela	98.994
41	tleng	97.905
42	jeso	97.857
43	ntse	93.265
44	robala	93.147
45	mokgatlho	92.883
46	mowa	86.582
47	nna	85.722
48	wena	81.657
49	pholo	80.714
50	teng	80.031
51	itse	78.64
52	raya	77.168
53	senatla	76.036
54	rraetsho	75.592
55	ralekgotla	72.16
56	burial	71.876
57	seabi	71.876
58	kwaletswe	67.288
59	moruti	66.894
60	christ	65.859
61	tsena	64.864
62	phuthego	64.433
63	malebogo	63.094
64	tlhodilwe	62.365
65	ululuIuuuu	61.607
66	hodisa	61.607
67	papa	60.058
68	sebui	59.819
69	baebele	59.764
70	nnyaa	59.736
71	wa	57.191
72	ko	56.166
73	kalo	55.268

74	ene	55.095
75	nae	54.952
76	tsaa	54.844
77	ha	53.71
78	tsile	53.587
79	buang	53.477
80	dumedisa	51.689
81	fela	51.471
82	rapeleng	51.339
83	tshegofatso	51.243
84	gonne	50.687
85	ra	50.603
86	matlhogonolo	50.052
87	three	50.052
88	go	49.754
89	dithaba	48.959
90	kgotleng	48.245
91	ira	48.194
92	mokgatlhong	47.949
93	lapeng	47.295
94	rata	47.028
95	tlaabo	46.927
96	khwaere	46.438
97	tshikinyega	45.944
98	na	45.591
99	masego	45.377
100	mmoloki	44.531

Most of the keywords are from the Christian religion. This is evident in the list with words such as: *Modimo* (3) (God), Amen (7), *morena* (15) (lord), *Keresete* (39) (Christ), *sefela* (40) (hymn), *Jeso* (42) (Jesus), *Christ* (60) *phuthego* (62) (congregation/a gathering), *Baebele* (69) (Bible), *rapeleng* (82) (pray), *tshegofatso* (83), *matlhogonolo* (86) and *masego* (99) (blessing(s)), and *mmoloki* (100) (saviour). From the funeral service words which stand out are mokgatlho (45) ((burial) society), burial (56), *sebui* (68) (a speaker), *kgotleng* (90) (the traditional meeting place where a meeting is held before burial), *ralekgotla* (55) (headman) *mokgatlhong* (92) (society) and *khwaere* (96) (choir). The funeral service as a formal gathering is characterised by words of respect such as *bagaetsho* (5) or *bagaetshong* (27) (fellow citizens), *Mme* (4) (Mrs), *Mma* (11) (mother of or Mrs.), *rra* (16) (Sir) and *rraetsho* (54) (Sir). The *Sidilega* program wishes the sick persons good health through the use of biblical scriptures. It therefore speaks of *pholo* (49) (healing) and *hodisa* (heal) (66).

The spoken subcorpus also comprises football commentary and other radio programs covering other types of sports. The keywords of this text are given in Table 50.

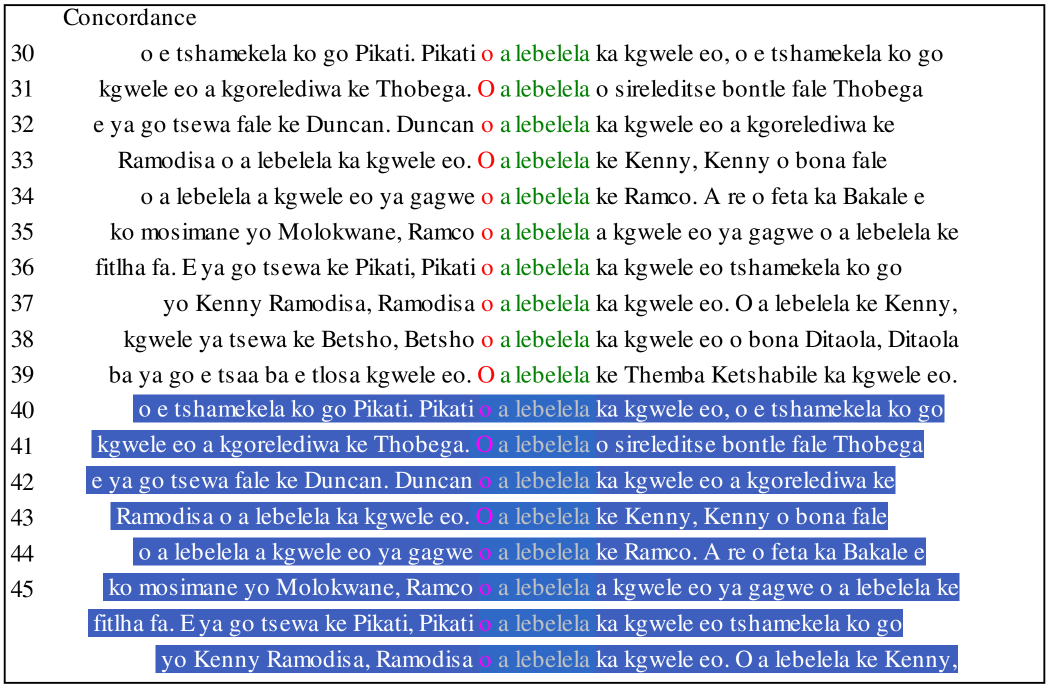
Table 50: Sport spoken text keywords

N	Keyword	Keyness
1	kgwele	1162.3
2	Ko	964.36
3	Ke	942.04
4	motshameko	562.01
5	team	559.96
6	setlhopha	461.49
7	mme	457.78
8	small	440.84
9	lebelela	436.12
10	mokatisi	421.29
11	coach	378.14

12	tshamekela	371.11
13	tsaya	353.65
14	e	344.49
15	ditaola	328.2
16	gore	325.92
17	le	308.5
18	league	305.12
19	tsewa	297.16
20	boy	295.42
21	viola	287.69
22	eo	286.04
23	tshameka	254.16
24	fale	251.46
25	tla	242.95
26	softball	239.85
27	ya	235.35
28	kadisa	226.86
29	pitcher	226.86
30	ka	224.71
31	lebelele	222.37
32	metshameko	220.67
33	pikati	202.82
34	wa	191.53
35	re	186.58
36	matius	181.02
37	ketshabile	181.02
38	lefela	179.72
39	gabaitsane	179.59
40	bakale	179.59
41	tatlhelo	179.59
42	mosimane	179.07
43	jwaneng	176.1
44	bona	175.03
45	fitileng	174.3
46	nako	174.23
47	tshutshu	170.14
48	na	168.23
49	ee	164.55
50	leng	162.25
51	jono	162.12
52	spears	160.69
53	setlhopheng	151.92
54	police	151.53
55	kgantele	148.74
56	kamoso	148.03
57	ntse	138.29
58	yone	137.87
59	tlosa	136.74
60	motshekgwa	134.32
61	wells	132.33
62	ecco	125.42
63	yole	125.03
64	masie	125
65	angels	125
66	blue	124.93
67	bokhutlo	122.16
68	tswela	117.96
69	netball	116.77
70	gunners	116.67
71	bokhutlong	114.06
72	tle	113.54
73	tournament	113.42
74	morapedi	113.21
75	satmos	112.52
76	go	111.02
77	tsatsing	108.1
78	gone	107.29
79	beke	106.52
80	chiko	106.39
81	player	106.39
82	kenny	106.39
83	fifa	105.4
84	okay	103.97
85	tshobega	103.97
86	motshamekong	101.73
87	lebe	101.18
88	duncan	97.821
89	modirelabangwe	97.105
90	themba	96.628
91	mabogo	96.217
92	bone	95.895
93	pitch	94.519
94	stopo	94.519
95	catcher	94.519
96	molokwane	94.519
97	shephi	94.519
98	teng	94.378
99	bokatisi	91.562
100	friday	89.968

The results of Table 50 are characterised by sport terms that refer to a variety of **games** such as *kgwele* (1) (ball or football), *softball* (26), and *netball* (69). Sport spoken text comprises football commentaries and interviews of sport personalities. Some of the terms that come up amongst the top 100 include amongst others names of **footballers**: *Ditaola* (15), *Viola* (21), *Kadisa* (28), *Pikati* (33), *Matius* (36), *Ketshabile* (37), *Bakale* (40), *Motshekgwa* (60) and many others. There are also **names of teams**: *Spears* (52), *Police* (54), *Wells* (61), *Ecco* (62), *Blue Angels* (66/65), *Gunners* (70), and *Satmos* (75). **Nouns and verbs common in sport** also rank high, amongst these being *kgwele* (1) (ball), *motshameko/metshameko* (4/32) (sport/game(s)), *setlhopha* (6) (team), *lebelela* (9) (watch), *mokatisi* (10) (coach), *coach* (11), *tshamekela* (12) (play for), *tsaya* (13) (take), *league* (18), *tshameka* (23), *pitcher* (29), *lebelela* (31) (watch), *lefela* (38) (zero), *tatlhelo* (41) (throw in), *nako* (46) (time), *tournament* (73), *player* (81). A variety of **sports** are represented as well amongst these being *softball* (26) and *netball* (69).

From the list above, there is an interesting use of certain words whose use appears to be unique to sport. Of note is the word *lebelela* which means 'to watch or observe'. One would expect the use of *lebelela* in the area of sport to have spectators as the subject and the game as the object of the verb. This however is not the case. The structure that we get in the concordance lines is that of *O a lebelela* (He is watching) followed either by *ka kgwele* (with the ball) or *ke Kenny Ramco (the name of a player).*



```
     Concordance
30           o e tshamekela ko go Pikati. Pikati o a lebelela ka kgwele eo, o e tshamekela ko go
31      kgwele eo a kgoreledisa ke Thobega. O a lebelela o sireleditse bontle fale Thobega
32      e ya go tsewa fale ke Duncan. Duncan o a lebelela ka kgwele eo a kgoreledisa ke
33         Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny, Kenny o bona fale
34           o a lebelela a kgwele eo ya gagwe o a lebelela ke Ramco. A re o feta ka Bakale e
35       ko mosimane yo Molokwane, Ramco o a lebelela a kgwele eo ya gagwe o a lebelela ke
36      fitlha fa. E ya go tsewa ke Pikati, Pikati o a lebelela ka kgwele eo tshamekela ko go
37             yo Kenny Ramodisa, Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny,
38        kgwele ya tsewa ke Betsho, Betsho o a lebelela ka kgwele eo o bona Ditaola, Ditaola
39       ba ya go e tsaa ba e tlosa kgwele eo. O a lebelela ke Themba Ketshabile ka kgwele eo.
40           o e tshamekela ko go Pikati. Pikati o a lebelela ka kgwele eo, o e tshamekela ko go
41      kgwele eo a kgoreledisa ke Thobega. O a lebelela o sireleditse bontle fale Thobega
42      e ya go tsewa fale ke Duncan. Duncan o a lebelela ka kgwele eo a kgoreledisa ke
43         Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny, Kenny o bona fale
44           o a lebelela a kgwele eo ya gagwe o a lebelela ke Ramco. A re o feta ka Bakale e
45       ko mosimane yo Molokwane, Ramco o a lebelela a kgwele eo ya gagwe o a lebelela ke
        fitlha fa. E ya go tsewa ke Pikati, Pikati o a lebelela ka kgwele eo tshamekela ko go
               yo Kenny Ramodisa, Ramodisa o a lebelela ka kgwele eo. O a lebelela ke Kenny,
```

In these instances when a player takes a football and looks to where he could pass it, the very act of searching for an unmarked player is expressed by the verb *lebelela*. This use of *lebelela* is unique to sport since the common use of *lebelela* is watch or watch over something.

Another word similar to *lebelela* which is unique to sport is the word *tsaya* (take). We first present its concordance lines

```
110   lhopha sele sa Police11. ba ya go e tsaya ba e tlosa kgwele eo. Ba ya
111   a re lebelele kgwele eo, ba ya go e tsaya ba e tlhoma. motsotso wa bom
112   wela kwa ntle kgwele eo. Ba ya go e tsaya ba e tlhoma ba etlosa go tlo
113   ousand and four . Kgwele ba ya go e tsaya ba e tlosa, motshameko e le
114   etsa go tswa kwa morago. Ba ya go e tsaya ba e emeletsa go tswa ka kwa
115    se kgobalo epe e masisi. Ba ya o e tsaya ba e tlosa setlhopa sa BDF11
116   osa e le tatlhelo BDF11. Ba ya go e tsaya ba e tlosa e le tatlhelo. Mo
117   ogo ya ga Oliver Pikati. Ba ya go e tsaya ba e tlhoma e le goal kick k
118   le kgwele e ntle. Kgwele ba ya go e tsaya ba e kolopa ba e emeletsa go
119   ousand and four . Kgwele ba ya go e tsaya ba e tlosa, motshameko e le
120   lopa Police11 kgwele eo. Ba ya go e tsaya ba e kolopa e kolopiwa fa le
121   a fela kontle kgwele eo. Ba ya go e tsaya ba e kolopa e le tatlhelo. B
```

To take something implies the use of hands. However in this instance this is not the case. It simply means being in possession of a ball.

Some of the words that collocate with *tsaya* are also unique to the genre of sports especially when referring to football. The words are *tlosa* (remove)*, kolopa* (throw at)*, tlhoma* (fix on the ground)*,* and *emeletsa* (raise up/lift up). *Kolopa* implies the use of hands to throw something at someone or something, however in this context it refers to kicking a ball into the air. *Emeletsa* is to raise something upright or on its feet. In the sports genre however it refers to setting a ball into flight. The use of these words indicates that words function differently in different contexts and their treatment in dictionaries need to reflect the different contexts in which they are used. For instance the treatment of the *tsaya* in Matumo (1993) may be improved by extracting concordance lines from a corpus and identifying collocates. *Tsaya* is entered in Matumo (1993: 426) thus:

> **tsaya** V.S. SIMP., take; take a wife; marry.

This entry can be improved this way:

> **tsaya** *v.* **1.** take with hands **2.** follow a path, choose a direction **3.** take a wife; marry **4.** be in possession of ◘ **tsaya botshelo**: take a life ◘ **tsaya dinopolo**: spy on someone. ◘ **tsaya ditaelo**: take orders. ◘ **tsaya ka motlhala**: follow. ◘ **tsaya dipilisi**: swallow pills. ◘ **tsaya dinopolo**: collects secrets ◘ **tsaya karolo:** take part. ◘ **tsaya ka letsogo**

la molema: illtreat; discriminate against. ◘ **tsaya kgakololo**: take advice. **tsaya kgato**: take a step. ◘ **tsaya lobaka**: take a long time. ◘ **tsaya mongwe/sengwe motlhofo**: undermine someone or something. ◘ **tsaya puso**: take over government. ◘ **tsaya phekelo e sele**: take a turn for the worst. ◘ **tsaya tshwetso**: take a decision. ◘ **tsaya nako**: take time. ◘ **tsaya motlhala**: copy an example from someone. ◘ **tsaya mosadi**: take a wife; marry. ◘ **tsaya mogote**: measure temperature. ◘ **tsaya matsapa**: put an effort. ◘ **tsaya tsia**: take someone or something seriously. ◘ **tsaya malatsi**: go on leave. ◘ **tsaya malebela**: copy something good. ◘ **tsaya maikarabelo**: take responsibility. ◘ **tsaya loeto**: take a trip. ◘ **tsaya maemo**: occupy a position. ◘ **tsaya setshwantsho**: take a picture. ◘ **tsaya sekgele**: win an award. ◘ **tsaya sebaka**: take time.

6.3 Conclusion to keyword analysis

In the preceding pages we have calculated keywords for Science and Technology, Politics, Poetry, Plays, Grammar, Arts and Culture, Religious and Spoken texts. We have presented the top 100 keywords from each genre or text type and shown that every text type contributes unique words. The findings support the position that for a corpus to represent the general language, it must be designed in such a way as to include a variety of text types from the language. The finding has been supported by keyword analysis which has revealed that the different text types generate different keywords that are particular to them.

The recognition that different text types contribute different words, should influence lexicographers compiling dictionaries on the basis of corpus evidence to pay particular attention to corpus design to ensure the broadest coverage possible of text types in a corpus. This is since the quality of retrieved information for lexicographic purposes depends on the information input at the stage of corpus construction.

Additionally, lexicographers could harness the power of keyword analysis and mark dictionary entries and senses on the basis of word variability. Many English dictionaries consistently mark frequencies (Kilgarriff, 1997; Summers, 1995); however much can be achieved by marking words or senses which rank high in a particular genre or text type. The challenge with raw frequency lists generated from a whole corpus is that they can push words which are high on the frequency analysis of a specific genre down on the frequency list of the whole corpus. The solution lies in an analysis similar to the ones conducted in this chapter which are genre based. As a result of keyword analysis in this chapter for instance, the words in Table 51 could therefore be entered in a dictionary and marked SPORT to indicate that they rank high in the keyword frequency analysis of sports terms.

Table 51: Possible SPORT candidates

English	English
kgwele	ball
motshameko	game
setlhopha	team
lebelela	watch/look
mokatisi	coach
tshamekela	play at
liki	league
tshameka	play
softball	softball
pitcher	pitcher
metshameko	games
tatlhelo	throw in

We illustrate the labelling with two dictionary entries *kgwele* and *setlhopha* from Matumo (1993).

> **kgwele** N. CL. 9N-, SING., any round object; commonly used to refer to a football.

> **setlhôpha** N. CL. 7 *se-,* SING OF *ditlhôpha*, a group; a company of people; a drove of animals.

The two entries could be improved with the SPORT label this way.

> **kgwele** *n.* [SPORT] **1.** a football. **2.** a ball.

> **setlhôpha** *n.* **1.** [SPORT] a sports team **2.** a group.

Marking entries as suggested will aid users and language learners in identifying the genre in which the word functions even before reading an illustrative sentence in an entry. Lexicographers could therefore device labels such as RELIGION, MUSIC, GRAMMAR, ARTS, NEWS, POLITICS, SCIENCE or LAW to make the dictionary more informative and user friendly.

In Chapter 7 we measure lexical density across text types at comparable token points. It will be established whether at comparable token points text types vary in lexical density and contribute different words. The diversity of lexical richness found in genres and domains is relevant for dictionary compilation since as argued before dictionaries should aim to be broad in their coverage of a language's lexicon.