Chapter 3

Corpus Lexicography

> It is important to avoid perfectionism in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like (Sinclair, 2004).

3.1 Introduction

At the turn of the century Kilgarriff epically observed: "The arrival of electronic text corpora is causing a revolution in lexicography" (Kilgarriff, 2000: 109). Kilgarriff's statement rings true for many lexicographic projects in various languages which are aided by the exploitation of corpora. Of note is the contribution of the British National Corpus (BNC) to the production of Longman dictionaries (Summers, 1995) and many others and the effect of the Bank of English on the COBUILD dictionaries (Sinclair, 1996; De Beaugrande, 1997 and Moon, 2007). We start by defining what a corpus is, how it is of benefit to lexicography and we discuss two basic ways in which corpora are usually exploited.

3.2 What is a corpus?

What a corpus is, is usually characterised differently by various scholars.

Leech (1991: 8) defines a corpus as "a sufficiently large body of naturally occurring data of the language to be investigated". On the other hand, Renouf (1987: 1) defines a corpus as "a collection of texts, of written or spoken word, which is stored and processed on computer for the purpose of linguistic research". Sinclair (2004) defines a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." McEnery and Wilson (1996: 24)

define a corpus as "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration."

What recurs in these definitions is that a corpus is a language sample, a collection of texts, or pieces of language text for linguistic research. Kilgarriff and Grefenstette (2003: 334) however would see Renouf, Sinclair and McEnery and Wilson's definitions as characterized by "a smuggling of values into the criterion of corpus-hood" and conflating questions: "What is a corpus?" with "what is a good corpus?" They argue that "a corpus comprising the complete works of Jane Austen is not a sample, nor is it representative of anything else" and they define a corpus as "a collection of texts," a definition they qualify thus: "when considered as an object of language or literary study."

From the above definitions various points may be noted.

1. Corpora are usually "sufficiently large" for the research they have been compiled for. They usually run into thousands or millions of words (e.g. the 100 million-word British National Corpus). What "sufficiently large" translates to in terms of number of words or size of file is however not clear.
2. Corpora are collections of running texts. They are not just lists of words but rather chunks of texts like chapters of books, entire books, or transcribed speech.
3. Corpora are compiled for some linguistic research. "With a corpus stored in a computer, it is easy to find, sort and count items, either as a basis for linguistic description or for addressing language-related issues and problems" (Kennedy, 1998: 11).
4. Because of their massive size, corpora are usually stored in computers because of their storage and processing power. Cowie (1999: 117) observes that "nothing less than a computer revolution had taken place in lexicography." Kirkness (2004: 56) argues that "computers can store and process quantities of textual data quite unmanageable by humans" (see also Biber et. al. 1998: 22). Computers aid in querying corpora in fast and sophisticated ways (Kilgarriff and Grefenstette, 2003: 333/334) and modern corpora analysis and storage are characterised by a dependency on computers. Computers are "good at recall,

people are good at precision; that is, computers are good at finding a large set of possibilities, people are good judges of which possibilities are appropriate" (Kilgarriff, 2003: 1). Several advantages of the corpus-based approach emanate from the use of computers which make it possible to identify and analyse complex patterns of language use, allowing the storage and analysis of a larger database of natural language possible. Computers also provide consistent, reliable analysis. They can also "be used interactively, allowing the human analyst to make difficult linguistic judgements while the computer takes care of record-keeping" (Biber et. al., 1998: 4).

In this thesis we follow Cavagliá (2005: 5) and limit our definition of corpora to "language corpora" and exclude other media such as pictures and sounds.

Biber et al. (1998: 4) list four essential characteristics of corpus-based research as follows:

i. It is empirical, analyzing the actual patterns of use in natural texts;
ii. It utilizes a large and principled collection of natural texts known as a "corpus" as the basis of analysis;
iii. It makes extensive use of computers for analysis, using both automatic and interactive techniques;
iv. It depends on both quantitative and qualitative analytical techniques.

Biber (1995: 32) also lists the advantages of corpus based analysis as including:

1. The adequate representation of naturally occurring discourse, including representative text samples from each register. Thus, corpus-based analyses can be used on long passages from each text, and multiple texts from each register.
2. The adequate representation of the range of register variation in a language; that is, analyses can be based on a sampling of texts of a large number of spoken and written registers.
3. The (semi-)automatic linguistic processing of texts enabling analyses of much wider scope than otherwise feasible. With computational processing, it is

feasible to entertain a comprehensive linguistic characterisation of a text, analysing a wide range of linguistic features. Further, once the software tools are developed for this type of analysis, it is possible to process all available online texts.

4. Greater reliability and accuracy for quantitative analyses of linguistic features; that is computers do not get bored or tired – they will reliably count a linguistic feature in the same way every time it is encountered.
5. The possibility of cumulative results and accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be based on the same corpus of texts, or additional corpora can be analysed using the same computational techniques. Such studies can verify the results of previous research, and findings will be comparable across studies, building a cumulative linguistic description of the language.

3.3 Web as corpus

One of the alternative methods of corpus compilation is the construction of corpora from the Web (Jones and Ghani, 2000; Ghani et al., 2001). The Web currently contains billions of words. Kilgarriff and (2003) report of 172 million network addresses in January 2003. Fletcher (2002) points out that the Web has over ten billion publicly-accessible online documents which provide a comprehensive coverage of the major languages and language varieties, and span virtually all content domains and written text types. This massive language data is particularly available in major European languages like English, French, Spanish, Italian, Dutch and German. Smaller languages like Setswana however are underrepresented. In Chapter 5 of this thesis we discuss how together with Kevin Scannell of the Department of Mathematics and Computer Science, Saint Louis University, we compiled about half a million Setswana tokens using a Web crawler and downloaded web text for adding into the Setswana corpus used in this study.

At a theoretical level the question to ask is whether Web language text qualifies as corpus data. To this question, Kilgarriff and Greffenstette (2003: 343) answer in the affirmative. They respond to the charge that the Web is not representative by arguing

that "the web is not representative of anything else. But nor are other corpora, in any well-understood sense." We discuss corpus representativeness in considerable detail in Chapter 4.

De Schryver (2002) discusses the Web as and for corpus in African languages. He demonstrates that although the Web is highly dominated by English text, African languages are represented on the Web and can benefit from exploiting Web corpus in language research, such as spell checking and checking for grammatical patterns. Languages such as Swahili, Amharic, Hausa, Silozi and Chinyanja, isiZulu and isiXhosa have been demonstrated to exist in good numbers online.

The Web provides a cheap route to corpus compilation. An illustration of this is Ghani et al., (2001 and 2001a) who report on the CorpusBuilder architecture, query-generation methods and language filters of downloading documents for minority languages from the Web. By minority they refer to languages which are in the minority on the Web not necessarily a language spoken by a few people. CorpusBuilder works by taking as initial input from the user two sets of documents, relevant and non-relevant. Given these documents, it uses a term selection method to select words from the relevant and non-relevant documents to be used as inclusion and exclusion terms for query, respectively. The query is sent to a search engine and the highest ranking document is retrieved. This results with a large collection of text within a short time.

Fletcher (2005) gives the following points as support for using Web text:

- **Freshness and spontaneity:** the content of compiled corpora ages quickly, while texts on contemporary issues and authentic examples of current, non-standard, or emerging language usage thrive online.
- **Completeness and scope:** existing corpora may lack a text genre or content domain of interest, or else may not provide sufficient examples of an expression or construction easily located online; some very productive contemporary genres (blogs, wikis, discussion forums…) exist only on the Net.

- **Linguistic diversity:** languages and language varieties for which no corpora have been compiled are found online.
- **Cost and convenience:** the Web is virtually free, and desktop computers to retrieve and process web-pages are available to researchers and students alike.
- **Representativeness:** as the proportion of information, communication and entertainment delivered via the web grows, language on and off the Web increasingly reflects and enriches our language.

Baroni and Ueyana (2006) isolate the following advantages and disadvantages of using Web corpora. First advantages:

- Size**.** The Web has large amounts of text. Text size is important in NLP. Disambiguation algorithm performs better when trained on a larger amount of data (see also Bindi et al., 1994).
- The Web allows fast and cheap construction of corpora in many languages for which no standard reference corpus such as the BNC is available to researchers.
- Web text can potentially contain a number of genres that are not present in traditional written sources such as blogs which generate vast amounts of spontaneously produced text.
- Web corpora tend to reflect more recent phases of a language than traditional corpora that are often subject to a certain lag between the time of production of the materials that end up in the corpus and the publication of the corpus.

They as well note the following disadvantages to Web corpora which are similar to those of any corpus built in a short time and with little resources.

- Web corpora are usually full of non-linguistic material and duplicated documents and duplicated text in different documents also referred to as ‘noise’.
- Since Web corpora are usually constructed with automated text mining methods, the researcher usually does not have full control over what ends up in the corpus, and cannot estimate the composition of the corpus.

- If a researcher plans to distribute a large Web corpus comprising millions of documents, (s)he will have a very hard time obtaining permission to use the documents from all the copyright holders.

Fletcher (2004) also finds the following disadvantages to Web-corpus data.

- It is difficult to establish authorship and provenance and to assess the reliability, representativeness and authoritativeness of texts since web-pages are typically anonymous and web server location is no certain guide to origin.
- Some sites have multilingual data.
- Other pages are authored by non-native speakers of varying competence, raising questions about language quality and influence of the source language.
- Certain longer prose text types predominate such as legal, journalistic, commercial and academic prose.
- Web text has no grammatical mark-up.

Some of the disadvantages listed by Fletcher are not unique to Web text. For instance, the argument that Web text has no grammatical mark-up is not limited to Web text since text from magazines and newspapers has no grammatical mark-up either.

Web text has been used for a variety of linguistic research. Amongst these is obtaining frequencies of bigrams unseen in a corpus (Keller et al., 2002). Shepherd and Watters (1998) propose what they term cybergenres and taxonomy of web-pages types and their evolution in an attempt to make sense of the structure of texts on the internet. Santini (2003) on the other hand proposes the development of computational methods to identify genres on the Web. Web text is therefore considered data that is suitable for linguistic analysis.

Electronic text corpora are valuable for language modelling in a variety of language technology applications such as speech recognition, optical character recognition, handwriting recognition, machine translation and spelling correction.

Kilgarriff (2001: 343) demonstrates that different researchers have used the web for a variety of research projects amongst these being:

1. As a source of language corpora for languages where electronic resources are in short supply.
2. As a source for bilingual parallel corpora.
3. To generate encyclopaedia entries.
4. For automatic distillation of lexical entries from empirical evidence.
5. In translation, translators when confronted with a rare term can find ample evidence of the term, its contexts, and associated vocabulary, through the simple use of a search engine.
6. The Web as a lexical resource, and as a source of test data, for Word Sense Disambiguation.
7. As a source for harvesting lists of named entities.

Similar benefits are discussed by Sharoff (2006).

The Web gives access to enormous quantities of text for free and it is still to be explored extensively in the study of Setswana. Chapter 5 reports on the use of a Web crawler to collect about half a million Setswana corpus from the Web. Below frequency profiling which will be used later in Chapter 6 and 7 is introduced.

3.4 Frequency profiling: frequency and type/token

3.4.1 Frequency counts

Frequency counts record the number of times each word occurs in a text. Sinclair (1991: 30) points out that "Anyone studying a text is likely to need to know how often each different word form occurs in it". This position is shared by Summers (1996: 261) that "all aspects of lexicography are influenced by frequency." Kilgarriff (1997: 135) furthermore notes that "A central fact about a word is how common it is. The more common it is, the more important it is to know it." Baroni (2006: 1) observes that "The frequency of words and other linguistic units play a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language." This

argument for frequency information is shared by Kilgarriff and Salkie who argue that:

> When a corpus is presented as a frequency list, much information is lost, but the tradeoff is an object that is susceptible to statistical processing. Word frequency lists are easy to generate, so measuring corpus similarity based on them will be viable in many circumstances where a more extensive analysis of the two corpora is not possible (Kilgarriff and Salkie, 1996: 121).

Baroni (2006: 3) observes that the data in a frequency list can be re-organized in two ways that are particularly useful to study word frequency distributions, namely as *rank/frequency profiles* and as *frequency spectra*. To obtain a rank/frequency profile, we simply replace the types in the frequency list with their frequency-based ranks, by assigning rank 1 to the most frequent type, rank 2 to the second most frequent word, etc. A frequency spectrum on the other hand is a list reporting how many word types in a frequency list have a certain frequency.

From our discussion in this section, it is clear that one of the basic corpus analyses is the frequency list, which reports a number of instances of each word type encountered in a corpus. Frequency counts therefore extract the different types of words, tokens or forms which make up a corpus.

A frequency list may be sorted in decreasing order from the highest ranking i.e. the most frequently used token down to hapax legomena (forms that occur only once in a given corpus) or vice versa. They are a "powerful tool in the lexicographer's arsenal of resources, allowing her to make informed linguistic decisions about how to frame the entry and analyse the lexical patterns associated with words in a more objective and consistent manner" (Summers, 1996: 266). We illustrate the use of frequency information in two English dictionaries *the Longman Dictionary of Contemporary English, 3rd edition* and *the Collins COBUILD Learner's Dictionary* in Section 3.5.

3.4.2 Type/token and word counts

In frequency analysis, there is a need to clarify what constitutes a *word* in a language and how words are counted. In linguistic literature the term *word* is defined in a

variety of ways. Some of these definitions while useful for theoretical linguistics, they are not very helpful in computational word counts. Finch (2000: 132) defines a word as "A unit of expression which native speakers intuitively recognise in both spoken and written language" and adds that "there is a certain indeterminacy about the definition of a word" (Finch, 2000: 132). Finch's definition is unhelpful in that "a unit of expression" could be anything from a word, a phrase, a clause or sentence. His definition also leaves the determination of what a word is to a speaker's intuition which may vary from one speaker to another. Aitchison (1992: 49) points out that "the best-known definition of a word is the one proposed by the American linguist Bloomfield who defined it as a minimum free form, that is, the smallest form that can occur by itself." She continues to argue that distinctions must be made between lexical items, syntactic words and phonological words. If we consider lexical items, the form such as *fly* represents at least two words:

fly N: an insect with two wings.
fly V: move through the air in a controlled manner.

The two lexical items have different syntactic forms associated with them. The insect could either be singular (*fly*) or plural (*flies*). The verb on the other hand could occur as *fly, flying, flies, flew, flown.*

Leech et al., (1982: 27) consider a word "delimited, for most purposes by a space (or punctuation mark other than a hyphen or apostrophe) on each side." However they also acknowledge that "the boundaries of words… are not always clear; e.g. we can write the sequence *piggy* + *bank* in three ways: *piggy bank*, *piggy-bank*, or *piggybank*."

In this study "a word is a minimal free form, the smallest unit that can exist on its own" (Dash and Chaudhuri, 2000: 189) and it is "delimited by a space…. on each side" (Leech et al., 1982: 27). From the brief discussions of what a word is it is clear that what a word is, is complex. Take for instance the following example:

> *His father will return from New York on Wednesday at 10am. Then they will stay in Pretoria for a week before buying a new house.*

One may observe that the above two sentences have 25 words: *His, father, will, return, from, New, York, on, Wednesday, at, 10am, Then, they, will, stay, in, Pretoria, for, a, week, before, buying, a, new* and *house.* Such a decision is arrived at by counting alpha-numeric characters delineated by spaces. Others may or may not consider whether digits and punctuation as words, a decision which will affect the shape of frequency distribution. Other decisions relate to token segmentation, whether it is *10 am* or *10am* or whether *New York* should be counted as a single token or two. Distinctions may also be made between upper and lower case forms. All these decisions will affect the number of elements counted. Distinctions between a word type and a token must be made so that there is no ambiguity and confusion concerning what is counted. Evert therefore emphasises that:

> For the purpose of obtaining frequency counts, it is essential to make a clear distinction between these two aspects: lexical items are called **types**, while their instances in a text are referred to as **tokens** (Evert, 2004: 33).

> A running word or a token is an arbitrary sequence (string) of letters delimited by spaces (Bergenholtz and Tarp, 1995: 34).

In a corpus of a 1,000 words the word form *kgomo* may occur 25 times. We say there are 25 tokens of *kgomo* which constitute a single word type. In this thesis, we follow Bergenholtz and Tarp (1995: 34) and assume words are separated by spaces and we count concords, grammatical words and numbers as distinct forms. A multi-word expression (MWE) will therefore be segmented where spaces exist and counted as multiple words. Words such as *gare ga bosigo* (midnight) or *bosigo gare* (midnight) will be counted as three and two distinct words respectively.

In African languages such as Setswana the matter of what a word is, is compounded by numerous concords in the language. While in English morphemes of verbs are suffixal and written conjunctively onto verbs, in Setswana the verbal prefixes (concords) are written disjunctively. This means that entities that are divided by white spaces are not always semantic words. Amongst these in Setswana are demonstrative concords, *lwa*, *jwa*; the quantitative concord, *ga*; relative concords, *sa*, *tsa*;

enumerative, *ga*, *wa*; nominal concords, *wa*, *ya*. The challenge could be illustrated with a text item such as *wa*. It could be a concord for first, second or third person singular of noun class 1 and noun class 1a. It could also be a concord for noun class three or eighteen. The use of the term 'word' in this thesis should therefore not be taken to mean that concords linguistically qualify as words and not morphemes. The use of the term word should rather be understood to mean graphical text items delineated by spaces. Such an approach will not be beyond criticism; however it does have some advantages to it; "the tradeoff is an object that is susceptible to statistical processing. (Kilgarriff and Salkie, 1996: 121).

3.5 Relevance of corpora to lexicography

Corpora are central to many lexicographic projects. De Schryver and Prinsloo (2000a: 292) note that "on the *macrostructural* level corpora provide crucial information for the creation of the lemma-sign list of dictionary, and on the *microstructural* level corpora enable lexicographers to tremendously enhance the accuracy of the dictionary articles themselves." On a macrostructural level they argue for lemmatised frequency lists and the need for lemma-sign distributions across sub-corpora to address typical macrostructural inconsistencies in many African-language dictionaries. They particularly expose the failure to include and treat commonly used words found in many other dictionaries as a lack of dependency on corpora. Their argument for the selection of lemma-lists on the basis of corpus frequency profiling is in line with Kilgarriff (1997: 136) who suggested the following processes were for improving the Longman Dictionary of Contemporary English, 3rd edition Summers, 1995) on the basis of frequency information:

- take a corpus
- extract a frequency list
- compare it with the dictionary to identify and rectify mistakes mismatches
- identify the one-, two-, and three- thousand cut-off points
- mark the corresponding dictionary entries accordingly

De Schryver and Prinsloo (2000: 298) argue not only that "frequency considerations

should determine the compilation of the lemma-sign list" but also that the lemma signs should be "in a sufficient variety of sources". This is what Leech et al. (2001: 17) refer to as word dispersion since 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." Scott (2004-2006: 123) refers to this phenomenon as consistency, that is, how consistent a word is used across a variety of texts of a corpus or subcorpus. Consistency analysis is calculated with every word frequency count in WordSmith Tools and is rendered in terms of the number of texts the word occurs in. This can be illustrated by considering the top 20 words in the Setswana corpus compiled for this study (see Chapter 5 for details). The top 20 tokens can be listed on the basis of decreasing frequency or on the basis of spread across texts (polytexty) as in Table 5 and Table 6. *Rank* refers to the position occupied by a token on a wordlist on the basis of its frequency. *Freq*. is the frequency of a word, or the number of times a word occurs in a corpus. *Texts* demonstrates the number of texts in which a word occurs. For instance "a" is ranked number 1 in Table 5 and it occurs 686,492 times in 3,055 texts.

Table 5: Top 20 words in the Setswana corpus ranked by word spread

Rank	Word	Freq.	Texts
1	a	686,492	3,055
2	go	418,088	3,000
3	e	413,176	3,016
4	le	358,736	2,977
5	o	336,417	2,990
6	ba	315,243	2,898
7	ka	290,557	2,956
8	ke	242,497	2,928
9	ya	228,511	2,959
10	mo	193,181	2,940
11	re	158,644	2,695
12	ga	149,529	2,851
13	fa	143,385	2,830
14	se	132,649	2,714
15	gore	125,686	2,828
16	di	124,651	2,807
17	ne	97,129	2,435
18	wa	94,822	2,803
19	tsa	92,885	2,772
20	sa	81,099	2,737

Table 6: Top 20 words in the Setswana corpus ranked by word spread

Rank	Word	Freq.	Texts
1	a	686,492	3,055
2	e	413,176	3,016
3	go	418,088	3,000
4	o	336,417	2,990
5	le	358,736	2,977
6	ya	228,511	2,959
7	ka	290,557	2,956
8	mo	193,181	2,940
9	ke	242,497	2,928
10	ba	315,243	2,898
11	ga	149,529	2,851
12	fa	143,385	2,830
13	gore	125,686	2,828
14	di	124,651	2,807
15	wa	94,822	2,803
16	tsa	92,885	2,772
17	sa	81,099	2,737
18	se	132,649	2,714
19	re	158,644	2,695
20	tse	69,238	2,640

Table 5 lists words on the basis of how frequent they are in the corpus. The most frequent word is ranked first, and the 11th frequent word, ranked 11, and so on. Table 6 on the other hand, lists words on the basis of spread, with a word found in the most number of texts ranked first. For instance, although *ba* is ranked 6th in terms of raw frequencies, it is ranked 10th in terms of spread. The rank of *ba* may be compared to that of *ya* which is ranked 9th in terms of raw frequency but is ranked in 6th in terms of spread.

Additionally, some of the words found in the raw frequency list do not make it into those listed in terms of spread. *ne* which occupies the 17th spot amongst the top 20 raw frequency list does not make it into the top 20 words on the basis of spread. Other words like *tse* which did not appear on the raw frequency list have been introduced into the list sorted by spread.

The two lists illustrate the fact that words that constitute a headword list should be abstracted on the basis of both raw frequency and word spread.

Corpora also have been used in the refinement of the dictionary microstructure, aiding in sense distinctions, the retrieval of typical collocations, frequent word clusters and the selection of authentic examples (De Schryver and Prinsloo, 2000a). De Schryver and Prinsloo (2000a) demonstrate that, particularly for African languages where this has been a weakness, corpora can tremendously aid to improve the quality of dictionary entries. Concordances, illustrative sentences in a dictionary and other linguistic information, which would be hard to generate through the use of non-corpus methods, are made readily available by the use of corpora and a corpus query system (CQS). For instance the sophisticated exploration of corpora by sketch engines (Kilgarriff and Rundell, 2002) has proved to be an efficient way of exploring the behaviour of words, the grammatical relations within which they participate, their collocational behaviour and thesaurus and "sketch differences" (which specify similarities and differences between near-synonyms) (Kilgarriff et al., 2004). A corpus therefore provides lexicographers with the information they need to compile authoritative descriptions of the vocabulary of a language. Lexicographers can retrieve the following from a corpus:

- **Statistical information**. From a corpus we can derive information about the relative frequency of different words or of the different grammatical constructions of the same word. This will reveal the inventory of the most common words which may be included as part of the dictionary's headword list, and the ones which are rare that may be left out of a dictionary. Thus Gomez (2002: 236) observes that frequency lists "enable lexicographers to take important decisions on which words a dictionary should include and which particular meanings". It is also possible to mark a word's frequency in dictionaries.

  We give examples of the marking of word frequencies in two English dictionaries: *the Longman Dictionary of Contemporary English, 3rd edition* (LDCE3) and *the Collins COBUILD Learner's Dictionary* (COBUILD).
  The LDCE3 was compiled using three corpora totalling over 135 million words: the 100 million words British national Corpus, the 30 million words Longman Lancaster Corpus and the 5 million word Longman Learner's Corpus. The dictionary marks the most frequent 6,000 words (3,000 entries from spoken transcribed text and another 3,000 from the written text) in the corpus on the page margins alongside entries. Spoken text is marked with S

and the written with W. The 3,000 words are further divided and marked by whether they are part of the top 1,000 spoken or written corpus (S1 or W1), the next thousand (S2 or W2) or the last thousand (S3 or W3) of the 3,000. For instance, the word **catch** is marked S1W1 to mean that it is part of the top 1 000 words in both spoken and written English. **Driver** on the other hand is marked S1W2 to mean that it is part of the 1,000 words of spoken English and it falls somewhere between the most frequent 2,000 and 3,000 words of written English. Such coding is essential since it guides a learner to words they are likely to meet and therefore need to learn.

Other frequency information in the dictionary, like meaning and homography, are not coded. The dictionary enters the most frequent meanings of a word first, and the less frequent ones later. For instance **chicken** has seven different meanings **1. ►BIRD◄ 2. ►MEAT◄ 3. ►SB WHO IS NOT BRAVE◄ 4. ►GAME◄ 5. which came first, chicken or egg? 6. a chicken and egg situation/problem/thing etc 7. your chickens have come to roost.** The meaning of *'bird'* is therefore more frequent in the language than *'somebody who is not brave'*.

Homographs are also shown in frequency order. The most common ones are entered and defined first while the less common ones are dealt with later. For example; **bound$_1$** (past tense of **bind**), **bound$_2$** (to be very likely to do…), **bound$_3$** (to run with a lot of energy) **bound$_4$** (noun, as in 'by leaps and bounds'). "by leaps and bounds" is rarer compared to **bound,** the past tense of **bind** in English. A learner would therefore be better off learning the past tense of **bind** before learning **bound** meaning "to run with lots of energy". Not only that, learners would be more likely to meet, in most texts, the most common meanings and if they look them up in a dictionary they would find them handled first, and not tucked in at the end. Such arrangement of senses is convenient since it ensures that words and meanings that students are likely to meet are arranged on the basis of their frequency.

The COBUILD gives frequency markers of entries to indicate how frequently they occur in the language. Instead of the S1 and W1 found in the LDCE3, they use a series of five diamonds ♦♦♦♦♦ in the extra column of the dictionary page. If all the diamonds are filled, then a word is one of the most frequent in the English language. The least frequent word has only one diamond filled ♦◊◊◊◊. There are nearly 700 entries representing 1,500

different forms which have five filled diamonds. The frequency of an entry includes that of its different forms (it is lemmatised), so that the frequency of the word *do* includes *does, doing, did,* and *done*. The next band of four filled diamonds ♦♦♦♦◊ covers over 1,000 entries which account for about 2,500 forms. Together with the five-filled diamonds band, the four filled diamonds words represent 75% of all common English usage. The 1,700 entries then represent essentially the core of the English language which is essential for a student to master. The next two bands of three black diamonds and two black diamonds ♦♦♦◊◊ cover a further 4,400 entries. The two filled diamond words ♦♦◊◊◊ include such words as *shuttle, shy, sickness, shrub, shrink, mounted, minimal, minus, midst,* and *soap*. Entries with a single black diamond ♦◊◊◊◊ represent the rare but important words which might have a restricted context of usage, they may be literary or words with specialized usage. The back matter of the dictionary comprises over 3,000 entries (Sinclair, 1996:1316-1322) accounting for nearly 10,000 forms. The decision to list them as part of the back matter is useful since students can assess their vocabulary power by simply reading the list and identifying those words that are unfamiliar to them and then finding their meanings in the dictionary. Teachers too can use the wordlist as a basis of class exercises to teach learners the core English vocabulary.

- **Alternative forms and spellings**. The corpus, if it is large enough, should present alternative spellings/forms of words (e.g. program/programme) and facilitate judgements as to which form should be used for the primary spelling. It will also reveal which forms are common enough to need to be entered as cross-references.

- **Semantic information**. From a corpus lexicographers can extract evidence for the word's different meanings and nuances. If the corpus is large enough, it would be able to show the most common senses of the word, and suggest the order in which such senses should be listed in a dictionary. It should also demonstrate common applications of the word (Biber et al., 1998).

- **Collocations**. From a corpus lexicographers can extract concordance lines to study the company that words keep. The computer concordance will reveal the most common collocations for individual words, by the words which tend to

come immediately before and after it in the concordance printout. It will assist describing the collocations which support particular senses, and their relative frequency would help to decide the order in which they should be listed in a dictionary. Collocation analysis will also lead to the isolation of idiomatic expressions (see Section 3.10).

- **Typical stylistic contextual information**. A large corpus with written and spoken data can provide examples from a variety of texts and sources, and thus indicate whether a particular word is current across the stylistic scale, or confined to spoken or written text only. In dictionaries such information may be turned into stylistic labels, such as "colloquial", or usage notes "found mostly in spoken interaction". Baugh et al. (1996: 44) argues that a corpus is critical in providing *context of use* information in the following areas: register (formal, informal, slang, taboo, taboo slang etc.), special context of use (specialised, medical, law, literary, poetic), language variety (American, British, and Australian), general context of use, e.g., speaker attitude (approving, disapproving).

- **Examples of actual use of the word**. Authentic examples from the corpus help to show the grammar and semantic functions of words, to remind L1 readers of the patterns of usage, and teach them to L2 readers. The dictionary-maker then has to decide whether such examples should be found for all words (including more technical ones), and for all senses of a word. The contents of the corpus example may therefore be needed to complement the definitions. Sinclair (1991: 39) argues that the “initial evidence should always be… from the observation of language in use”.

Baugh et al. (1996: 44) additionally argues that a corpus is beneficial to dictionary compilation in that it demonstrates the typical subjects and/or objects of a verb (see Section 4.3 of Chapter 4) and reveals encyclopaedic information of a word.

The contribution of a corpus to the dictionary making process has been discussed extensively in lexicographic literature by Béjoint (2000: 97), Sinclair (1987) and Sinclair (1991) who demonstrate the crucial nature of a corpus to the dictionary

compilation process.

3.6 Some pre-electronic frequency studies

Studies of frequency lists culled from corpora are not a recent occurrence and predate corpus computational developments. Kennedy (1998: 13-19) reports on pre-electronic corpora before the 1960's in five main fields of scholarship which we summarise below:

1. **Biblical and literary studies:** From at least the 18th century the Bible as a corpus has been used to generate lists and concordances to show that the Bible parts were factually consistent with each other. An example of such work is Cruden's concordance of 1736.

2. **Lexicography:** Corpus lexicographic work may be traced to 17th century Samuel Johnson's large corpus of sentences from writers to illustrate meanings and uses of English words (Sinclair 1991: 40). The compilation of the *Oxford English Dictionary* (OED) by James Murray and associates was also corpus-based; with over 2000 readers collecting millions of citations to illustrate word usage. In America, Noah Webster compiled *An American Dictionary of the English Language* in 1828 with the help of citation slips comprising millions of words (Kennedy, 1998: 14).

3. **Dialect studies:** Nineteenth century linguists compiled corpora to explore lexical variation in the choice of words for particular concepts (Kennedy, 1998: 14/15).

4. **Language education studies:** Thorndike (1921) compiled a 4.5 million-word corpus from 41 sources and generated a frequency list to aid curricula materials for teaching. J.W. Kaeding with the aid of assistants developed an 11 million word German corpus to gather statistical information of German words and letters to improve the training of stenographers.

5. **Grammatical:** Other corpora have been compiled to be used as sources of descriptive grammars. Among these is the work of Jespersen (1909-49).

3.7 Electronic-corpora studies

The first electronic corpus was the one million *Brown University Standard Corpus of Present-Day American English* commonly known as the Brown Corpus by Francis and Kučera (1964). It comprised 500 samples of 2000 words of continuous written English.

A similar corpus to the Brown Corpus, the Lancaster-Oslo/Bergen (LOB) Corpus, was compiled in the late 70s to study British English (Johansson and Hofland, 1989) through frequency analysis. Recently mark-up and word frequency studies have been done on the 100 million-word BNC (Leech et. al., 2001 and Rayson et al., 2002). These detailed studies do not only list alphabetical and rank frequency lists of the whole corpus, but include frequency lists of spoken versus written parts of the corpus, and unlemmatized frequency lists of spoken and written parts of the BNC. The studies investigate the frequency lists of the demographically sampled and context governed part of the spoken BNC.

3.7.1 An example of frequency profiling

Below the value of frequency profiling is illustrated by studying words which are characteristic of a particular genre. We look at the sports and business text. For our experiment parts of about a million tokens of the *Mokgosi* newspaper text are used. *Mokgosi* was a Botswana newspaper that wrote exclusively in the Setswana language. It closed down in 2005. The *Mokgosi* newspaper text is divided into five categories, the number of tokens given in brackets: Arts & Culture (476,523), News (1,426,223), Letters (502,729), Sport (289,205) and Business (247,246). For this part we are only interested in Sport (289,205) and Business (247,246) subcorpora from which we generate frequency counts using WordSmith Tools version 4.0 (Scott, 2004-2006). From our results we give the top 100 words for each including functional words and then offer the results again of the top 100 words, with functional words excluded.

In Table 7 below *N* is the number a word occupies in the list in terms of its frequency, this is the same as its rank. *Freq*. is the frequency of a word, or the number of times a word occurs in a corpus.

Table 7: Top 100 Mokgosi sport tokens with functional words

N	Word	Freq.
1	A	11,596
2	E	10,999
3	o	10,175
4	go	6,682
5	ba	6,566
6	le	6,129
7	ka	4,397
8	ya	3,835
9	mo	2,569
10	fa	2,315
11	ke	2,253
12	se	2,162
13	re	2,117
14	gore	2,085
15	di	1,890
16	ga	1,850
17	ne	1,742
18	sa	1,724
19	wa	1,600
20	kwa	1,559
21	tsa	1,238
22	tse	1,177
23	la	1,120
24	tla	1,118
25	bo	893
26	mme	748
27	bone	739
28	fela	721
29	setlhopha	649
30	jaaka	552
31	nna	532
32	batshameki	520
33	jwa	489
34	motshameko	473
35	na	464
36	yo	461
37	ngwaga	454
38	aforika	445
39	jalo	429
40	neng	406
41	morago	402
42	metshameko	367
43	ditlhopha	357
44	kgaisanyo	356
45	gagwe	348
46	dira	328
47	lekgotla	324
48	madi	318
49	botswana	310
50	thata	283
51	lefatshe	279
52	borwa	275
53	ene	267
54	ntse	267
55	pele	266
56	tswa	264
57	mokgosi	258
58	bona	256
59	kgwele	247
60	teng	247
61	batho	243
62	nako	243
63	mafatshe	242
64	bangwe	235
65	bobedi	235
66	ntlha	232
67	mongwe	231
68	masome	225
69	sentle	222
70	tlhalositse	221
71	eo	210
72	yone	200
73	setse	197
74	kgwedi	189
75	simolola	189
76	dipitse	188
77	motho	188
78	gone	186
79	kgaisano	180
80	mono	180
81	gape	179
82	tshameka	178
83	dinao	176
84	rona	176
85	bile	169
86	jaanong	169
87	setshaba	168
88	komiti	165
89	tsenelela	160
90	itse	159
91	mabedi	158
92	lesome	154
93	tota	154
94	tshwanetse	152
95	sena	151
96	bomme	149
97	santse	148
98	tiro	148
99	tlhalosa	147
100	batla	145

Table 7 reveals that *a, e, o, go, ba, le, ka, ya, mo, fa*, are the top ten most frequent words in the sport subcorpus. These words are members of the closed word classes (also known as function or grammatical words) which include classes such as concords, pronouns, and numerals (Leech et al., 1982). At least 35% of the words in

Table 7 are functional words. We find that the first 28 words are all functional words. It is common to most frequency lists to have functional words at the top of frequency lists.

Therefore in determining the frequency of the most frequent tokens in a corpus it may be attractive to remove the functional words from the list. Since functional words are usually the most frequent in corpora, they may in certain cases not provide critically comparative information between lists. Removing them from wordlists and remaining with content words (open-class words) may aid lexical comparison between lists in certain cases. This argument is not new. Gomez has argued before for English analysis that:

> The main problem with this information (frequency list of functional words) is that the use of raw frequencies highlights the very common words such as *the, of, in etc.,* despite the fact that their comparatively high frequencies of occurrence are unlikely to provide conclusive evidence of any specifically used vocabulary in any sublanguage (or corpus). These are words that, on the basis of frequency of occurrence alone, would be found to occur within most sublanguages, and it can perhaps be read more usefully if the purely grammatical words (close-word items) are discarded (Gomez, 2002: 239).

The argument is therefore that the top 100 words would be read informatively if the functional words were discarded from the list. Their removal would reveal content words that could define a genre and provide comparative information. We therefore removed functional words from the *Mokgosi* Sport wordlist's top 100 tokens. The top 100 words excluding functional words are:

Table 8: Mokgosi sport list's top 100 tokens without functional words

N	Word	Freq.	8	neng	406	16	madi	318
1	Setlhopha	649	9	morago	402	17	botswana	310
2	nna	532	10	metshameko	367	18	thata	283
3	batshameki	520	11	ditlhopha	357	19	lefatshe	279
4	motshameko	473	12	kgaisanyo	356	20	borwa	275
5	ngwaga	454	13	gagwe	348	21	ntse	267
6	aforika	445	14	dira	328	22	pele	266
7	jalo	429	15	lekgotla	324	23	tswa	264

24	mokgosi	258
25	bona	256
26	kgwele	247
27	teng	247
28	batho	243
29	nako	243
30	mafatshe	242
31	bobedi	235
32	ntlha	232
33	mongwe	231
34	masome	225
35	sentle	222
36	tlhalositse	221
37	setse	197
38	kgwedi	189
39	simolola	189
40	dipitse	188
41	motho	188
42	kgaisano	180
43	mono	180
44	gape	179
45	tshameka	178
46	dinao	176
47	bile	169
48	jaanong	169
49	setšhaba	168

50	komiti	165
51	tsenelela	160
52	itse	159
53	mabedi	158
54	lesome	154
55	Tota	154
56	tshwanetse	152
57	sena	151
58	bomme	149
59	santse	148
60	tiro	148
61	tlhalosa	147
62	batla	145
63	seka	145
64	tshwana	145
65	nngwe	144
66	fetileng	143
67	dilo	142
68	fenya	141
69	jang	140
70	batswana	139
71	gompieno	139
72	rre	138
73	kgang	137
74	motshameki	137
75	sengwe	135

76	dingwe	133
77	tlang	133
78	gae	132
79	nno	130
80	basimane	129
81	maemo	129
82	ise	127
83	bontsi	126
84	liki	125
85	metshamekong	120
86	tsile	120
87	jaana	119
88	nne	118
89	pedi	116
90	tshameko	114
91	morule	113
92	bfa	112
93	gaborone	112
94	mathata	112
95	tsaya	112
96	tsena	112
97	bnsc	110
98	boletse	110
99	tlase	110
100	dikgaisanyo	109

The list of content words reveals clearly the genre of sport through the use of the following words *setlhopha* (team) (1), *batshameki* (players) (3), *motshameko* (game) (4), *metshameko* (games) (10), *ditlhopha* (teams) (11), *kgaisanyo* (competition) (12), *madi* (money) (16), *nako* (time) (29), *simolola* (start) (39), *Dipitse* (Zebras – a nickname for the Botswana football team) (40), *kgaisano* (competition) (42), *setšhaba* (nation) (49), *basimane* (boys) (80), *liki* (league) (84), *tshameko* (play, noun) (90), *pedi* (two) (89), *BFA* (Botswana Football Association) (92).

The frequency list has helped isolate the most common words that are characteristic of the genre purely on the basis of their frequency. However other words in the top 100 wordlist are not distinctive to the genre. Such words include *tshwana* (same as), *tlhalosa* (explain), *tota* (truly), *fetileng* (past), *ise* (has not), *boletse* (told/said), *jalo* (like that), *mathata* (problems), *Morule* (December), *jaanong* (now), *dilo* (things), *maemo* (positions), *tsaya* (take), *batla* (want/seek), and a few others. This is not surprising since the top 100 words are raw frequency outputs and are not isolated on any measure that isolates words which are typical to, or stand out in, a text.

3.8 Keyword analysis

A much more precise method of identifying words particular to a genre is through the calculation of keyness which isolates words which are "key" to a corpus or subcorpus since these are useful in characterising a text or genre. We will implement our calculations by using a KeyWord tool which is part of WordSmith Tools version 4 (Scott, 2004-2006). The program has been used previously successfully for comparing corpora (Berber-Sardinha, 2000; Scott, 1997, Xian and McEnery, 2005).

To conduct the calculations, two corpora or subcorpora are required: one large another small. The large one is used as a reference file, while the small one is the study corpus, the one we are interested in studying. A reference corpus has 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). Two wordlists are generated from the two corpora. The aim is to find out which words characterise the text that is analysed. Keyness is "calculated by comparing the frequency of each word in the wordlist of the text you're interested in with the frequency of the same word in the reference wordlist" (Scott, 2004-2006: 92). The result is a list of keywords, or words whose frequencies are statistically higher in the study corpus than in the reference corpus. These are known as *positive keywords*. The software also identifies words whose frequencies are statistically lower in the study corpus. These are called *negative keywords*. In this study it is the positive keywords that we are interested in i.e. words occurring with a higher frequency than expected.

For this experiment the study corpus is the *Mokgosi* Sport section with 289,205 tokens and we compare it against our reference corpus, for which we will use the *Mokgosi* News section text which has 1,426,223 tokens. We provide the results below of only the top 100 most frequent tokens.

Table 9: Mokgosi top 100 sports keywords

N	**Keyword**	2	setlhopha	4	kgaisanyo
1	batshameki	3	motshameko	5	ditlhopha

6	metshameko
7	kgwele
8	kgaisano
9	dipitse
10	aforika
11	motshameki
12	tshameka
13	liki
14	bfa
15	bnsc
16	tshameko
17	mokatisi
18	kgaisanong
19	thenese
20	metshamekong
21	tla
22	dikgaisanyo
23	sofotebolo
24	nno
25	tsenelela
26	sejana
27	kgaisanyong
28	morule
29	zone
30	dinno
31	popa
32	basimane
33	setlhopheng
34	ngwaga
35	volleyball
36	motshamekong
37	karate

38	ketlogetswe
39	ikatisa
40	ikatiso
41	dinao
42	boramabole
43	dikgwele
44	oosi
45	bdf
46	batabogi
47	nosa
48	notwane
49	mabelo
50	mokatise
51	veselin
52	mokganedi
53	rollers
54	bobedi
55	fenya
56	lefela
57	tunisia
58	lebaleng
59	iponela
60	Diolimpiki
61	mabole
62	dipetsana
63	fighters
64	dietsele
65	fani
66	morocco
67	motshwara
68	liking
69	marumo

70	keatlholetswe
71	kemoeng
72	molefhe
73	soobolo
74	luza
75	tlhaodi
76	netebolo
77	borwa
78	nigeria
79	komiti
80	cosafa
81	bakatise
82	motsotsong
83	botsamaise
84	ditlhopheng
85	tafic
86	mono
87	championships
88	mafolofolo
89	kutlwano
90	libya
91	dikgaisano
92	ikatisong
93	molwantwa
94	karolong
95	dikgaisanyong
96	fifa
97	phenyo
98	kirikete
99	rugby
100	tshamekile

The above *Mokgosi* sports 100 keywords offer us a better streamlined list of terms that are key in the genre of sports. This list is more precise than a list generated on the basis of frequency. It isolates those terms which are only key to the genre from the corpus. The list includes names of **teams** like *Dipitse* (9), *Popa* (31), *BDF* (45), *Notwane* (48), *Rollers* (53), *Tunisia* (57), *Fighters* (63), *Nigeria* (78), *Tafic* (85), *Mafolofolo* (88), *Kutlwano* (89); names of different **sports/games**: *kgwele* (football) (7), *thenese* (tennis) (19), *sofotebolo*/*soobolo* (softball) (23/73), volleyball (35), karate (37), *netebolo* (netball) (76), *kirikete* (cricket) (98), rugby (99); names of **sports associations and organisations**: BFA (14), BNSC (15), COSAFA (80), FIFA (96); names of **sport personalities**: (*Tom*) *Ketlogetswe* (the name of a sports journalist) (38), *Veselin* (the name of the former Botswana national football team coach) (51), *Mokganedi* (the name of a sports journalist) (52), *Marumo* (the name of a footballer)

(69), *Kemoeng* (the name of Botswana National Sports Council Chairperson) (71), *Molefhe* (the name of an athlete) (72), *Tlhaodi* (The assistant chairperson of Botswana Tennis Association) (75), *Molwantwa* (the name of a footballer) (93), *Luza* (the name of a boxer) (74), and many others. The list also includes **sport verbs** amongst these being *tshamekile* (played) (100), *tsenelela* (take part in/attend) (25), *ikatisa* (train) (39), *fenya* (win) (55), *iponela* (got/won) (59). The list includes **sport positions** amongst these being *komiti* (committee) (79), *botsamaise* (leadership) (83), *motshameki/batshameki* (player/players) (11/1), *mokatise* (coach) (50). That the top 100 most key tokens contain data from diverse games including names of officials and players suggests that keyness analysis is crucial for isolating data that is particular to a genre.

3.9 Business keywords

The keyword analysis experiment was repeated to extract business keywords. For this experiment the study corpus is the *Mokgosi* Business section with 247,246 tokens and it is compared against our reference corpus which is again the *Mokgosi* News section text which has 1,426,223 tokens.

Table 10: Mokgosi business keywords

N	keyword
1	ditlhaeletsanyo
2	kgwebo
3	ndlovu
4	thapelo
5	ditshupo
6	banka
7	penrich
8	kompone
9	boripana
10	dikgwebo
11	itholo
12	dibanka
13	bagwebi
14	letlole
15	dikompone
16	bojanala
17	dibonto
18	bedia
19	bobs
20	agoa
21	aforika
22	peepa
23	dithoto
24	mafatshe
25	koporase
26	yuropa
27	khemikhali
28	lekalana
29	air
30	letlalo
31	madi
32	funeral
33	kgwebong
34	provida
35	thulaganyo
36	dipesente
37	ditlamelo
38	ditlhotlhwa
39	koafatsa
40	botlhole
41	boccim
42	lenaneo
43	itsholelo
44	mookamedi
45	reka
46	alafasegeng
47	beci
48	kelebogile
49	tlhalosa
50	foxcroft
51	hemilwe
52	hemiwa
53	sacu
54	lenchwe
55	diselula
56	tsogwane
57	sekoloto
58	lefhenya
59	rialo

60	difofane
61	lenaneong
62	kotsi
63	mhama
64	mmaraka
65	barclays
66	tlhong
67	makasine
68	mokgopha
69	peugeot
70	privatisation
71	siwawa
72	solofelwa
73	thekiso

74	dithentara
75	diphatsa
76	taolo
77	bota
78	galeforolwe
79	dikoketso
80	jab
81	tlhotlhwa
82	ceda
83	allan
84	batlamedi
85	diresiti
86	golden
87	kgokagano

88	nshakazhogwe
89	okacom
90	ppadb
91	privatization
92	thema
93	boranyane
94	mono
95	mabenkele
96	diaparo
97	boherabongwe
98	orange
99	papadisanyo
100	tlaabo

The business text in the *Mokgosi* newspaper is written by the business journalist, Thapelo Ndlovu. His surname and first name occupy position 3 and 4 respectively in the above list. The list also includes names of the following **companies, businesses and organizations**: *Penrich* (7), *BEDIA* (18), *BOBS* (19), *AGOA* (20), *BOCCIM* (41), *BECI* (47), *SACU* (53), *Barclays* (65), *Peugeot* (69), *BOTA* (77), *JAB* (80), *Allan* (for Allan Gray) (83), *OKACOM* (89), *Orange* (98); **business nouns** *ditlhaeletsanyo* (communications) (1), *kgwebo* (business) (2), *kompone* (company) (8), *letlole* (saving) (14), *dikompone* (companies) (15), *dibonto* (bonds) (17), *dithoto* (goods) (23), koporase (corporation) (25), *lekalana* (department/sector) (28), *madi* (money) (31), *itsholelo* (economy) (43), *sekoloto* (debt) (57), *mhama* (sector) (63), *dithekiso* (sales) (73), *dithentara* (tenders) (74), *diresiti* (receipts) (85), *boranyane* (technology) (93), *mabentlele* (shops) (95), *diaparo* (clothes) (96), *papadisanyo* (trade) (99) and many other terms; **business personalities**: *Thapelo Ndlovu* (business journalist) (2/3), *Kelebogile* (*Rantsetse*, the name of a local entrepreneur) (48), (*Slumber*) *Tsogwane* (Assistant Minister of Finance and Development Planning) (56), *(chief) Lenchwe* (54), (*Kagiso*) *Lefhenya* (young entrepreneur) (58), (*Tshidi*) *Tlhong* (chairperson of Junior Achievement Botswana) (66), *Mokgopha* (the name of a cobbler) (68), (Anthony) *Siwawa* (chairperson of CVF) (71), (*Joshua*) *Galeforolwe* (chairman of PEEPA) (78), (*Ishmael*) *Nshakazhogwe* (chairperson of Zambezi Motors) (88).

The relevance of these lists is to be evidence for the power of frequency profiling in lexical analysis. Frequency counts assist in the identification of most significant words on the basis of their frequency. Keyword analysis aids the retrieval of different genre-specific terms. When such analyses are repeated on texts from different genres and text

types, we would end up with keyword lists from different genres of a language, which could be combined to provide a broad vocabulary of such a language. For lexicography, the challenge with compiling a corpus with different text types of different sizes and then studying the raw frequency lists of the entire corpus together is that such an approach may obscure the keywords in various corpus components since certain keywords from a particular text type may be pushed lower in the frequency list and therefore risk exclusion. Studying different genres in isolation can aid in ensuring that different genres are represented and reflected in the dictionary headword list. The results may also be crucial in aiding marking entries as frequent in certain genres.

In Chapter 6 and 7 we use frequency profiling to measure lexical density and isolate keywords from a variety of genres in the Setswana corpus. Such experiments will be aimed at proving that subcorpora are characterised by different words and that their inclusion in a corpus to make a broad-coverage corpus is essential for an accurate representation of linguistic diversity in corpora. A broad-coverage corpus is a source of diverse linguistic wealth for dictionary compilation.

3.10 Concordance

Another way of studying words in a corpus is by studying a specific word in context in some detail in terms of co-texts to the left and to its right. This is achieved by generating a key word in context (KWIC) often referred to as concordance lines. “A concordance is an index of the surface word forms in a text. It is a collection of the occurrences of a word form, each in its own textual environment” (Dash and Chaudhuri, 2000: 190). A concordance reveals the company kept by a word, its collocates, and thereby reveal meanings and usages which are hard to dig up through mental recall. We illustrate this below through the example of the word *pelo* (heart*).*

Figure 1: Concordance results of the word *pelo*

```
o ka kgopolo ya gore Morwadi o tlaa wela pelo. A mo gaupanya. ka legofl fa ga re
 ngwatiaka, O se tshoge bono wa ka, Wela pelo ga o seitaodi re Use rotlhe, O sek
gang, o tla e rola morago o sena go wela pelo. '/'r~ a emelela, o b~ua a le esi)
a ka seatla. "0 sale sentle, moratiwa wa pelo ya me. Ga ke itse gore ke~ tla go
se o lela jalo? Ke a go rata moratiwa wa pelo ya me." Fa a sa ntse a e phimola,
ne a ithuta ona. "Gomotsega, moratiwa wa pelo, ya me." Mosele a didimala, mme go
 rebe la Mokwena, a buledisa moratiwa wa pelo ya gagwe. O ne a tsamaya ka bonya,
a. : Ke go reile ka re o seka wa utlwisa pelo botlhoko tlhe rra! 0 a itse gore b
```

```
g mo matlhong a gago. O se ka wa utlwisa pelo ya gago botlhoko ka nna, ke swetse
 gatlhisa thata., Ba,utlwil ba mo tswela pelo tota.. ,I, Mmaago Molebi a mo roma
be, a di phailela kwa, a re ba mo tswela pelo. Le ene Pule tota tsala ya gagwe y
hegelwa ke moratiwa, e seng go mo tswela pelo kgotsa go mo tlhoafalela. Seno se
sadi yo montle, mme phokojwe a mo tswela pelo. Phokojwe a leka maano a le mantis
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
ro ya gagwe. NtsVwa Mosetsanyana ya tswa pelo, ya metsa mathe a keletso. Saitsan
swe ke marago a tanka e nngwe. A tshwara pelo monnamogolo wa batho mme a ntse a
gotla-tshekelo. Mmaagwe Sereri a tshwara pelo ya gagwe, a bua ka tidimalo le bad
e motlha mongwe lokwalo lo tla tshikinya pelo ya ga Mmatheebe. A kwalela kgarebe
ologeletsweng morwadi wa we la tshikinya pelo ya gagwe gore a bale ka mabogo a a
nala kwa  Naledi a tla a ratile go tlola pelo. Mogoma e re ntlhomane a feta fa M
o ngwega Uncle Boot 0 ne a rata go tlola pelo. Mmaagwe ene, a ipega fa a ne a le
r ." Bikibiki a re, "Ngwana yo o tlhomola pelo. Lefatshe le mo itaya ka ntlha ya
 re ruri rre Rapitso wa batho o tlhomola pelo ka tshenyo e e leng ka mo supamake
 le bobotlana. Motho wa tsona o tlhomola pelo. Keikepetse o ntse jalo mo teramen
ka go tena Ontefile. Ketshedile a tlhapa pelo ka o ne a sa ntse a senka leano la
a. Bofelo a mo tsepela leitlho. A tlhapa pelo gonne fa go rata Bofelo, tsotlhe d
ng. jaanong Morwadi a nametsega a tlhapa pelo. A tsaya tlhogo a e latsa mo sehub
otlhapelo a ngwana wa mpa. "Nnake, tiisa pelo. O sa ntse o na le mogomotsi e bon
 bosigo. Tlogela go nna legatlapa. Tiisa pelo. Ga o sa tlhola o le mosimane. Ka
a gagamatsa thamo ya gagwe, a thatafatsa pelo ya gagwe, mo a bileng a se ka a bo
2Mme le ka sebaka seo Farao a thatafatsa pelo ya gagwe gape, a se ka a naya mora
wa gee." "Ke mang?" A botsa a swegaswega pelo. Ngwananyana a bolela fa ene a bon
a ke matlhagatlhaga, e bile a swegaswega pelo. O ne a batla go balela kwa pele.
```

The word *pelo*'s English equivalent is *heart* "a hollow muscular organ that pumps the blood through the circulatory system by rhythmic contraction and dilation" (Pearsall, 1998: 847). In the concordance lines above, *pelo* taken together with its collocates, is rarely used to convey the meaning of the physical heart. In the first line, *wela pelo* literally means "have your heart fall down" meaning "be at peace or settled." *Moratiwa wa pelo* (the loved one of the heart) is equivalent to "sweet heart" or "beloved". *Tshwara pelo* (handle or hold the heart) means "be in control of your emotions." It is through inspecting collocates that we can uncover proverbs, compounds, idioms, sayings, phrasal verbs and different multi-word expressions. Such structures could then be entered in dictionaries as sub-entries. Through the use of computer programs or concordance software it is relatively easy to get a list of all the cooccurences of a particular word in context and see all the meanings associated with the word (Biber et. al., 1998: 27). The concordance lines above reveal the different subtle meanings associated with the word *pelo*. From such a study of concordance lines, we have extracted possible subentries of the *pelo* headword. We have been able to extract 84 possible sub-entries (see Appendix 1); below we give only 10 of these.

Table 11: Corpus derived possible subentries of *pelo* entry

Collocates	Literal translation	Meaning
Ama pelo	Touch the heart	Hurt someone
Balabala ka pelo	Speak too much by heart	Talk aloud to yourself; absent minded
Baya pelo	Put the heart	Relax
Beta pelo	Suffocate the heart	Persevere
Betwa ke pelo	Be choked by the heart	Be very angry
Bofa pelo	Tie the heart	Restrain yourself
Bolawa ke pelo	Be killed by the heart	Desiring something
Bolwetse jwa pelo	The disease of the heart	Heart attack
Bona pelo	See the heart	See one's intentions or their thoughts
Bua ka pelo	Speak with the heart	To be troubled to the extent that you speak to yourself

The phenomenon of idiomaticity when considering a word and its collocates is not unique to the word *pelo* in Setswana. Words like *molomo* (mouth), *nko* (nose), *monwana* (finger), *kgomo* (cow) and *mpa* (stomach) all display similar characteristics. Such idiomatic expressions can enrich dictionary entries as subentries.

Table 12: Corpus derived possible subentries of *mpa* entry

Collocates	Literal translation	Meaning
Bana ba mpa	Children of a stomach	Relatives
Bipa mpa ka mabele	Cover the stomach with breasts	withhold bad information to protect a relative or friend
Gare ga mpa ya bosigo	In the centre of the belly of the night	In the middle of the night
Gare ga mpa ya lefatshe	In the centre of the stomach of the world	In the middle of nowhere
Gare ga mpa ya naga	In the centre of the belly of the wilderness	In the middle of nowhere
Mpa ya sebete	The belly of the liver	Flat on the stomach
Mpa e tuka molelo	A belly burning fire	Filled stomach
Go ja ka mpa tsoopedi	To eat with two stomachs	To eat until the stomach is full
Ntsha mpa	Take out a stomach	Commit abortion
Imelwa ke mpa	Be overladen with a belly	Have a full stomach

Table 13: Corpus derived possible subentries of *molomo* entry

Collocates	Literal translation	Meaning
Bolwetsi jwa tlhako le molomo	The disease of hoof and mouth	Foot and mouth disease
Itoma molomo wa tlase	Bite the lower mouth	Be determined
Itshwara molomo	Hold/touch a mouth	Be shocked
Ntsha ka molomo	Release with the moth	Speak

Pula molomo	That which opens the mouth	Money paid before someone speaks in lobola negotiations
Pipa-molomo	That which covers the mouth	A bribe
Rwala molomo	Carry the mouth on your head	To be angry and tight lipped
Roka molomo	Sew the mouth	Remain quiet
Tswa molomo	Grow mouth	Speak
Tlhoka molomo	Lack a mouth	Have nothing to say

Table 14: Corpus derived possible subentries of *lonao/dinao* entry

Collocates	**Literal translation**	**Meaning**
Apaya ka lonao	Cook with a foot	Avoid cooking and instead eat in other people's homes
Goga dinao	Drag feet	Move slowly
Fodisa dinao	Cool feet	Have a rest
Motsamaya ka dinao	One who walks with feet	A pedestrian
Ngotla dinao	Reduce feet	Reduce walking pace
Tlhatlosa dinao	Raise feet	Increase walking pace
Baya lonao	Put a foot	Be in a place
Tsholetsa dinao	Lift feet	Increase walking pace
Kgwele ya dinao	A ball of feet	Football
Tsosa dinao	Wake up feet	Increase walking pace/hurry up
Tiisa dinao	Strengthen feet	Increase walking pace

Table 15: Corpus derived possible subentries of matlho entry

Collocates	**Literal translation**	**Meaning**
Bula matlho	Open eyes	Educate/make aware/open eyes
Diga matlho	Drop eyes	Look down
Digalase tsa matlho	Glasses of the eyes	Spectacles/sunglasses
Latlhela matlho	Throw eyes	Look briefly
Matlho a phage a lebane	The eyes of a wild cat face to face	Face to face
Kala matlho	Measure eyes	Confuse
Tlodisa matlho	Make eyes jump	Overlook someone or something
Kgarakgaratsha matlho	Make eyes move from one place to another	Look from one place to another
Tlhatlosa matlho	Raise eyes	Look up
Tlhaetsa matlho	Shorten eyes from	Despise someone

Setswana dictionaries have attempted to include subentries based on the idiomaticity of collocates. However some of these have been few because of a lack of sufficient corpus evidence. Below we give examples of the treatment of *molomo* in different Setswana dictionaries.

Brown (1925: 210)

> **Molomo**, n., pl. melomo, A mouth (outside); a beak of a bird; a foreskin. *Kgwedi ea molomo*, the first month of the Sechuana year; the month of eating fruits. *Go cwa molomo*, to open the mouth, in speaking.

Kgasa (1976: 71)

> **molomo(me)** kgôrô e dijô di yang mo 'ganong ka yônê.

Kgasa and Tsonope's (1998: 171).

> **mo•lomo** TTT ln./3. me-. phatlha e e tswalwang ke dipounama tse pedi e go tsenngwang dijô ka yônê go ya ko mpeng le go bua. ♠ *molomo o tlola noka e tletse = motho o kgôna go bua dilô tse di ntsi tse a ka di dirang mme ntswa a se ka ke a kgôna*

Matumo (1993: 260)

> **molomo**, N. CL, 3 *mo-*. SING. OF *melomo*, a mouth; lip; a beak of a bird; an opening, as a tube, piping or tunnel; a foreskin. ID. EXPR., *go tswa molomo,* to open the mouth in speaking. PROV., *sejô sennye ga se fete molomo.*

Snyman et al (1990) does not enter *molomo*.

All the dictionary treatments of the *molomo* entry above are deficient and will benefit tremendously from the use of corpus evidence. For instance, the Matumo (1993) definition may be revised in the following way:

> **molomo**, *n.* 1. mouth 2. a lip. 3. a beak. 4. an object opening, as that of a bottle. ◘ **bolwetsi jwa tlhako le molomo**: foot and mouth disease. ◘ **itoma molomo wa tlase**: be determined. ◘ **itshwara molomo:** be shocked. ◘ **ntsha ka molomo**: speak; express an opinion; express a view. ◘ **pula molomo**: money paid before someone speaks in lobola negotiations. ◘ **pipa molomo**: a bribe. ◘ **rwala molomo**: be angry and tight lipped. ◘ **roka molomo**: remain quiet. ◘ **tswa molomo**: speak; say something; contribute; express an opinion. ◘ **tlhoka molomo**: Have nothing to say; be dumbstruck; be rendered speechless. ◘ **molomo o tlola noka e tletse**: it is easy for someone claim that they can achieve what they cannot do.

In the revised entry above ◘ is used to mark a subentry. Thus the study of collocations can enrich the dictionary entries. Thus we conclude this section by illustrating how dictionary entries for *pelo, mpa, matlho, molomo* and *lonao* could be enriched on the

basis of information in Tables 11, 12, 13, 14 and 15 derived from a corpus. We compare the proposed entries with entries from Matumo (1993). Matumo (1993: 306/7) enters twenty subentries for *pelo*. We have shown that over eighty sub-entries could be extracted from a corpus (see Appendix 1).

Matumo (1993: 276)

> **mpa** N. CL, 9Ø-, SING. OF *dimpa*, a belly; a stomach. ID. EXPR. *mpa ya lentswê*, the middle of a hill; *mpa ya lonao*, the sole of a foot. PROV., *seboba re bata sa mokwatla sa mpa re a mpampetsa.*

Matumo's *mpa* entry might be improved in this way:

> **mpa** *n.* a belly; a stomach. ◘ **bana ba mpa**: relatives ◘ **bipa mpa ka mabele**: withhold bad information to protect a relative or friend ◘ **gare ga mpa ya bosigo**: in the middle of the night ◘ **gare ga mpa ya lefatshe/naga**: In the middle of nowhere ◘ **mpa ya sebete**: flat on the stomach ◘ **mpa e tuka molelo**: with a full stomach ◘ **go ja ka mpa tsoopedi**: to eat until the stomach is full ◘ **ntsha (senya) mpa**: commit abortion ◘ **imelwa ke mpa**: have a full stomach.

Matumo (1993: 212)

> **lonaô** N. CL. 11 *lo-*, SING OF *dinaô*, a foot. ID EXPR, *go baba lonaô*.

Matumo's *lonao* entry might be improved in this way:

> **lonao** *n.* a foot ◘ **apaya ka lonao**: avoid cooking and instead eat in other people's homes ◘ **goga dinao**: move slowly ◘ **fodisa dinao**: have a rest ◘ **motsamaya ka dinao**: a pedestrian ◘ **ngotla dinao**: reduce walking pace ◘ **tlhatlosa dinao**: increase walking pace ◘ **baya lonao**: be in a place ◘ **tsholetsa dinao**: increase walking pace ◘ **kgwele ya dinao**: football ◘ **tsosa dinao**: increase walking pace ◘ **tiisa dinao**: increase walking pace.

Matumo (1993: 232)

> **matlhô** N. CL. 6 *ma-*, PL OF CL. *leitlhô*; *maitlhô* is still used in a few areas, eyes.

Matumo *matlho* entry might be improved in this way:

> **matlhô** *n.* eyes. ◘ **bula matlhô**: educate, make aware, enlighten ◘ **diga matlhô***:* look down ◘ **digalase tsa matlhô**: spectacles, sunglasses ◘ **latlhêla matlhô**: look briefly ◘ **matlhô a phagê a lebane**: face to face ◘ **kala matlhô**: confuse ◘ **tlodisa matlhô**: overlook someone or something ◘ **kgarakgaratsha matlhô**: look from one place to another ◘ **tlhatlosa matlhô**: look up.

By proposing improvements for the dictionary entries, we hope to have illustrated the power of corpus evidence and concordance lines. However corpus generated collocations and frequency lists have not always been used to inform the complexity of a dictionary entry. Other methods have been explored which we discuss briefly below.

3.11 A review of existing methods of headword list identification

We have argued how a corpus can be used as a source of headword and subentries. However, a corpus has not and is not always used by lexicographers for dictionary compilation. In this section we review different methods which have been used by lexicographers to identify headword lists for dictionary compilation. For ages, lexicographers battled with ways and means of producing authentic and reliable reflections of the lexicon for different languages. Most of these lexicographers depended on their ability to remember words that existed in the languages under study, something that Prinsloo and De Schryver (2000: 4) call entering "words as they cross the compiler's way" and Kilgarriff (2000: 109) call "the lexicographer's intuition". Others on the other hand, in the Oxford tradition, depended on readers, who searched texts for word occurrences and submitted citations of words for entry into the dictionary. The readers' contribution, for many years, made the OED (Oxford English Dictionary) the most comprehensive lexicographic work of the English language. Developments in lexicography, later proved that readers were not reliable sources of dictionary material since they did not only take too long to process data, but they also could not accurately deliver information on matters of frequency across texts and genres to aid decisions on what to include and exclude (Summers, 1995, Sinclair, 1996 and Kilgarriff, 1997: 135).

Since the revolutionary COBUILD research using corpora evidence of 1981 (Sinclair, 1987 and Sinclair, 1991 and Moon, 2007: 159) there has been a rapid increase of dictionary projects that depend on corpora. The earlier Birmingham school of corpus lexicography adhered religiously to a corpus as a source of dictionary evidence. They argued that corpora were the sole sources of lemmatisation, frequency information wordlists and authentic examples (Fox, 1987: 138/9). If a word was not in a corpus it was not recognised as legitimate dictionary material.

However as corpus lexicography develops, the focus does not lie on corpus output exclusively, but more crucially, on corpus design and composition since they determine corpus output. Matters of representativeness, balance and genre coverage become more urgent to both theoretical and practical lexicographers (For a detailed discussion of this matter, refer to Chapter 4 of this thesis). Researchers want to know the nature of texts that form a corpus and in what proportions they stand to each other (Kilgarriff, 1996). Therefore the greatest challenge lies not so much in what we get from a corpus, but rather in its construction, for it is what goes into a corpus that determines what can be extracted from it.

For the remainder of this chapter we focus the discussion on the headword list identification. We begin by sketching the English tradition of headword list identification. We then proceed to discuss the non-corpus approaches to dictionary compilation and end by looking at corpora use in Setswana dictionary compilation.

3.12 A historical perspective of headword lists

The earlier English dictionary compilations where characterised by two phases; the Latin-English dictionaries and the dictionaries compiled by direct borrowing from literary works, especially technical terms appended to learned vernacular publications of the time.

The need to list words may be traced to the seventh and eighth centuries when "priests and scholars, glossing Latin manuscripts, compiled lists of difficult words to help readers unfamiliar with Latin" (Wells, 1973: 13). These lists grew longer and were subsequently presented in alphabetical order for easy access. They developed into what became Latin-English, English-Latin bilingual dictionaries. This laid the foundation for what could be termed dictionaries of "hard words" tradition of the 16th century.

For African languages it was the European explorers (Naden, 1993; Lichtenstein, 1928-30) and missionaries (Moffat, 1826) in the 1800s who recorded languages either out of curiosity or for Bible translation purposes.

By the 17th Century, in the English tradition, it was clear that lexicographers depended on reading many books for compiling wordlists as Bailey (1736) states in his preface that he depended on "the reading of a very large number of authors…" (quoted in Wells, 1973: 21). This approach was also adopted and developed by Samuel Johnson who "added a new empiricism, a wide ranging program of reading diverse sources" (op. cit. 21). Bailey compiled a 40,000 entry dictionary, the *Universal Etymological English Dictionary* (1721) (see Osselton, 1983).

The problem of collecting words has posed great difficulties to English lexicographers for a long time, as it currently does to Setswana lexicographers, especially for those who were interested in not merely copying other dictionaries.

> But to *collect* the *words* of our language was a task of greater difficulty: The efficiency of dictionaries was immediately apparent; and when they were exhausted, what was yet wanting must be sought by fortuitous and unguided excursions into books, and gleaned as industry should, or chance should offer it, in the boundless chaos of a living speech. My search, however, has been either skilful or lucky; for I have much augmented the vocabulary (Johnson, 1963: 10).

Johnson here points to three sources he used for his dictionary: other dictionaries, books, and "living speech". Some of the texts were from as diverse sources as science, technical dictionaries and philosophical writings.

> On the terms of art I have received as could be found either in books of science or technical dictionaries; and have often inserted, from philosophical writers, words which are supported perhaps only by a single authority, and which being not admitted into general use, stand yet as candidates or probationers, and must depend for their adoption on the suffrage of futurity (op. cit.).

Although Johnson's method was not completely corpus-based, it does point to a dependence on diverse sources of texts and spoken language for dictionary material.

3.13 Non-corpus dependant methods of dictionary compilation

Although in this thesis we argue for corpus methodology in dictionary compilation, there are still many practising lexicographers who use other strategies in compiling dictionaries. Ooi (1998: 47/48) identifies two different ways in which lexical or lexicographic evidence is derived for inclusion in dictionaries. These are lexical introspection and casual citation. By lexical introspection is meant the lexicographer's linguistic introspection, the words he can remember. Casual citation refers to "when the lexical behaviour of one's family members, friends, or strangers is observed and recorded" (op. cit. 48). In this instance lexicographic evidence is based on the people a lexicographer comes into contact with.

Other dictionaries have been used as sources of lexicographic evidence as seen in Section 3.12. Zgusta argues that "An important source of information can be found in other dictionaries of the language in question, if there are any" (Zgusta, 1971: 239). For the compilation of old dictionaries and even in some modern dictionary compilation practices, lexicographers have copied other dictionaries. The practice predates Johnson's dictionary. Thus Wells notes that,

> Lexicographers have traditionally borrowed quite freely from preceding dictionaries, sometimes plagiarizing with a free hand… [and] more often existing dictionaries were consulted and synthesized with other sources such as spelling books and technical glossaries (Wells, 1973: 21).

However Svensén warns against this practice noting:

> …there is a type of evidence which can be misleading, and curious enough is to be found in dictionaries. Straightforward errors in one dictionary may, when this in turn is used as a source for another dictionary, come to be regarded as authentic linguistic productions. This gives rise to 'ghost words', i.e. words which do not actually exist (Svensén, 1993: 41).

But Svensén's caution is not new; Johnson saw the weaknesses of his predecessors in including words which could not be accounted for anywhere in written texts or speech, and decided to omit them from his dictionary. Bailey, Ainsworth and Philips are lexicographers who had published dictionaries before Johnson.

> Many words yet stand supported only by the name of Bailey, Ainsworth, Philips, or the contracted *Dict.* for *Dictionaries* subjoined: of these I am not always certain that they are read in any book but the works of lexicographers. Of such I have omitted… (Johnson, 1963: 12/13).

Another method for gathering text dictionary compilation is the use of semantic domains developed by Ronald Moe of the SIL (Summer Institute of Linguistics).

3.14 Semantic domains

Moe (2001) of SIL (Summer Institute of Linguistics) proposes a method of semantic domains to be used for the collection of words. He argues that the methodology is particularly attractive for minority languages, most of which have none or few written texts, or no corpora. His argument is that the methodology is 100 times faster than collecting words without a structure. He argues that 12,000 words have been collected in a few weeks through what is effectively a simple methodology but one which is able to produce a massive classified dictionary and thesaurus.

Moe analysed domain classification of words as suggested by Murdock et al. (1987), Roget's (1958 and 1985 editions) and Louw and Nida (1989) and found them inadequate for eliciting vocabulary. What Moe attempts to compile is "a universal list of semantic domains" (Moe, 2001: 151) which field lexicographers could use to prompt native speakers to think of words in their language. However, semantic domains have greater relevance than mere elicitation of mother tongue speakers' words. "It could be used to collect words, it could serve to classify a dictionary, and it could aid in semantic investigation" (Moe, 2001: 152). Underlying this system is a mental approach to the lexicon; that words are all linked together in the mind in a gigantic multi-dimensional web of relationships which cluster around a central nexus (Moe, 2001: 4). The mental

lexicon is not alphabetical but words cluster around key concepts and it is these concepts that Moe calls semantic domains (Moe, 2003: 216). It is therefore his argument that related words should be collected at the same time. To guide field workers, Moe phrases domains as questions as in the following for the domain 'sing":

What words refer to singing? sing, serenade, warble, yodel, burst into song
What words refer to singing without using words? hum, whistle

These series of questions are central to what Moe calls the Dictionary Development Process (DDP) which he used in Uganda in training lexicographers in collecting Lunyole (a Bantu language) words. The DDP has 1,700 domains each with 8-10 questions which could elicit over 10 words per domain which means that the dictionary would have at least 17,000 entries.

Moe's semantic domain approach is relevant for the kind of context for which it has been constructed – minority languages with a very limited written tradition. It will prove very useful for individuals who gather words in rural areas and communities with languages with none or limited written tradition, where there may be a lack of written texts and technology to capture and process oral data of such languages. The semantic domain approach may also be used to augment a wordlist compiled from what could be perceived as an imbalanced corpus as a result of a lack of text from a specific genre.

While the semantic domain method may be used for gathering words of lesser-known languages of the world with limited or no written tradition, for languages with a large body of written texts this tedious task may not prove essential since huge corpora could be compiled which could be queried cheaply in various sophisticated ways. The semantic domain method of lexical collection does not provide any frequency information. Rather it in effect enters words into a dictionary as they are remembered by respondents.

For the purposes of this thesis we favour data from a well designed corpus. By a well designed corpus we refer to a corpus that comprises samples of language varieties from a language of interest. Central to the use of corpora is that linguistic information that goes into making a dictionary 'must be authentic, that is to say it must include only such

linguistic occurrences as actually exist… the lexicographer must find evidence for it in independent sources" (Svensén, 1993: 40).

3.15 Corpus lexicography and Setswana dictionaries

Of the entire Setswana dictionary compilations discussed in Chapter 2 of this thesis, it is Kgasa and Tsonope (1995) who report the use of a corpus in the compilation of their dictionary. They point out that:

> Re dirisitse tsa maranyane a dikhompiutara go tlhotlha le go runa mafoko a feta dikete di le makgolo a mabedi le masome a matlhano (250, 000) mo dikwalong di le mmalwa; ra tloga ra a oketsa ka mafoko a mediriso-puo e e faphegileng jaaka maina a dinaledi, dinonyane, mebala le matshwao a diphologolo, ditlhare, ditlhaga le dimela tse dingwe, ditiro tsa Setswana jalojalo (Kgasa and Tsonope 1995: v-vi).

> We have used computer technology to analyse a corpus of more than 250, 000 words from a compilation of several books; we then added special terms such as names of stars, birds, animal colour terms, names of trees, grasses and other plants, and terms particular to the Setswana culture etc (translation mine).

Kgasa and Tsonope are pioneers of corpus use in Setswana lexicography, particularly in Botswana. Their compilation of a corpus of a quarter of a million Setswana tokens in 1990 was an enormous achievement in an environment where Setswana language texts were not readily available.

While their corpus compilation is commendable, little is known about its structure and quality since their corpus construction process is not documented in any publication that we are aware of save for what we have quoted above from the introduction to the dictionary.

It is also clear that Kgasa and Tsonope did not sample any spoken language for their corpus. This is not surprising since the compilation of spoken corpora is both tedious and expensive. However a lack of transcribed speech in corpora has lead to deficiencies which have been observed in the literature. We had argued against such an imbalance:

> Such an imbalance raises questions relating to the composition and balance of the corpus. This is so since speech is the primary channel of human communication and exists in abundance compared to written text. (Otlogetswe, 2004: 194).

We have also argued (Otlogetswe, 2006: 150-153) that the exclusion of spoken text results in a loss of instances of borrowings from Afrikaans and English which are not usually accepted in the written Setswana form by many publishers.

Other dictionaries like *Dikišinare ya Setswana English Afrikaans Dictionary Woordeboek* are purely introspective in their approach. No wonder the compilers say:

> The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compiling of a dictionary. This can take place simply because the lexicographer had not encountered such words. We can only hope that there are not too many examples of this kind (Snyman et al., 1990: preface).

Matumo's (1998) is an updated version of Brown (1925). It does not make any claim of corpus use.

3.16 Conclusion

In this chapter we have demonstrated how frequency lists aid in the determination of words commonly used in a corpus. We have argued that such information can assist the lexicographer to compile headword lists. We have illustrated how the exclusion of functional words from the frequency list may reveal clearly the words that are typical of a corpus or subcorpus. While frequency lists have proved to be useful in the identification

of genre-specific words they have been found limited. We have shown how keyword analysis assists in the isolation of genre specific words which could form part of a headword list. Keyword analysis is also significant in genre identification in sociolinguistics and lexicography. We use Keyword analysis in detail in Chapter 6. In addition to frequency analysis, we have argued that a corpus can be exploited through concordance lines. Concordance lines are significant in revealing words which occur in the company of others. They unearth collocation, idiomatic expressions, phrasal verbs and various multi-word expressions.