Chapter 4

Issues in corpus design for lexicography

> One of the main issues addressed here, though, is whether general language studies must be based on a corpus that is register-diversified as well as large (Biber, 1993: 220).

4.1 Introduction

In the previous chapter we have considered what a corpus is and a variety of ways in which it is exploited for different ends. In this chapter we look at issues which arise in corpus design, particularly as they relate to the area of lexicography. Corpus design is relevant to this thesis since at the heart of this thesis is the argument that corpus design and compilation determine the quality of what could be extracted from it. The area of corpus design is broad and an attempt will be made to cover some of its most fundamental matters. Atkins et al. (1992) present a detailed discussion on corpus design through a panoramic overview of corpus design including practical stages of compiling a corpus including text selection and mark-up; the problems of defining a population of texts to be sampled; the types of corpora and their various uses. Some of the issues they raise will be investigated in considerable detail in this chapter.

As the use of computer-based text corpora has become increasingly important for research in natural language processing, lexicography, and descriptive linguistics, issues relating to corpus design have also assumed central importance (Biber, 1993: 219). Therefore a "corpus which is designed to constitute a representative sample of a defined language type" (Atkins et al., 1992: 2) has become increasingly attractive. Samples may be divided into two broad categories of written and spoken text. Written text refers to such written products as books, novels, magazines and letters. Spoken text refers to transcribed speech from meetings, lectures, telephone conversation, interviews or debates. These two broad categories are characterized by variability.

It is a linguistic truism that language is characterized by varieties (Fromkin and Rodman, 1998: 400-404). These varieties may be sociolects or social dialects, that is, linguistic varieties on the basis of facts such as socioeconomic status, gender, ethnic grouping, age, occupation and others (Southerland and Katamba, 1996: 540). There are also regional varieties; distinct linguistic varieties which characterise people from a certain geographic area. Linguistic varieties may also be perceived from the perspective of functional speech varieties also known as registers which characterise language on the basis of whether it is casual, formal, technical and other characteristics (Hudson, 2000: 452).

The recognition of a lack of linguistic uniformity in speech communities has relevance to corpus design since it means that "…due to the importance and systematicity of the linguistic differences among registers, diversified corpora representing a broad range of register variation are required as the basis for general language studies" (Biber, 1993: 219). We therefore differ with some proponents of very large corpora who have "suggested that size can compensate for a lack of diversity – that if a corpus is large enough, it will represent the range of linguistic patterns in a language, even though it represents only a restricted range of registers" (Biber, 1993: 220).

The design of corpora for lexicography comprising a diversity of texts raises multiple issues which are the subject of this chapter. These matters include amongst others: balance and representativeness, corpus size, corpus annotation, sample size and spoken language in a corpus. We begin by the subject of balance and representativeness.

4.2 Balance and representativeness

Biber (1995: 130/131) notes that in the area of social sciences, issues of representativeness are dealt with under the rubric of *external validity*, which refers to the extent to which it is possible to generalize from a sample to a larger target population. However there are two kinds of error that can threaten external validity: *random error* and *bias error*. *Random error* occurs when the sample is not large enough to accurately estimate the true population; *bias error* occurs when the selection of a sample is systematically different from the target population. Random error can be minimised by increasing the sample size, and this is why large text corpora are important. Bias error on the other hand refers to the sampling of only a part of a population to the exclusion or limited inclusion of other parts of the population. In contrast, bias error *cannot* be reduced by increasing the sample size, because it reflects systematic restrictions in selection. That is, regardless of corpus size, a corpus that is systematically selected from a single register or limited varieties cannot be taken to represent the patterns of variation in an entire population. Rather, in order to make global generalizations about variation in a language, corpora representing the full range of registers are required. Bias error therefore has to be addressed by broadening the representation of linguistic variability in a corpus.

The matter of balance and representativeness is one of the greatest areas of contestation in corpus design and compilation. On one hand, there are those who argue that a language can be sampled in its varieties to form a corpus that can be taken as a representative sample of the whole language. For instance, Renouf points out that:

> When constructing a text corpus, one seeks to make selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested (Renouf, 1987: 2).

There are those who argue that since we can never know all the varieties of a language and researchers possess no facts about the amount of spoken or written text that exist in real life, there is no way anyone can claim to compile a corpus that can be representative of the whole language. We explore both arguments.

4.2.1 Proponents of balance and representativeness

Biber et. al. (1998: 246) state that a corpus is not just a collection of texts, but at the heart of corpus design and construction is an attempt at creating a representative sample of a language or parts of a language that can be studied. Representativeness here according to Biber should be understood to mean "the extent to which a sample includes the full range of variability in a population" (Biber, 1994: 378). The "full range of variability" here refers to the range of text types and of linguistic distributions in a language. Therefore this means the object that is represented needs to be well understood by a compiler since "an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from the population" (Biber, 1994: 378). This position is similar to the one held by Renouf (1987: 2) who argues that "The first step towards this aim [constructing a corpus] is to define the whole of which the corpus is to be a sample." Biber et al. show that one of the problems in sampling is characterising the language to be sampled. However one of the limitations of attempting to characterise the language is that "we do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all variation in texts" (Biber et al., 1998: 246).

While the full varieties of a language may be unknown, there are other simpler cases where the whole text to be analysed may be finite and known as in the case of the total works of Shakespeare or the whole Bible text (Renouf, 1987: 2). Kilgarriff and Grefenstette however contend that, "A corpus comprising the complete published works of Jane Austen is not a sample, nor is it representative of anything else" (Kilgarriff and Grefenstette, 2003: 334) since it is the complete works of a specific writer.

Language can also be sampled proportionally. Such sampling will translate to highly used varieties sampled in greater proportions compared to rarely occurring ones. This will mean that since speech is used more in human communication compared to written language, corpora would have higher levels of spoken language compared to written language. A corpus designed in this manner approximates Biber's rough estimates:

> A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing (Biber, 1994: 386).

Such a corpus could be considered representative only in that it approximates how different varieties are used in a language. Biber (1994) however argues that proportional representativeness is not interesting for linguistic research. What is interesting however is "language samples that are representative in the sense that they include the full range of linguistic variation existing in a language." The major weakness with proportional sampling of language (i.e. both produced and received language), Biber has argued, is that even if it could be achieved, it would result with relatively homogenous corpora. This is because most texts in such corpora would be from conversation therefore having similar linguistic characteristics, since speech is proportionally greater than written language. A proportional sample may therefore not include texts from registers which are rarely read by the public such as legal and medical documents (cf. Burnard, 1995).

Biber et al. (1998: 89) therefore point out that a "key aspect of corpus design for most studies, then is including the range of linguistic variation that exists in a language, not the proportions of variations." They argue for *stratified* sampling which involves cataloguing the different categories of texts that exist in a language and sampling each of them, instead of *proportional* sampling which tries to compile proportions of language varieties that people use and receive.

Their argument is therefore that corpus language variability must approximate the linguistic variability of a speech community under study or if it does not, corpus limitations should be acknowledged. Biber (1995: 27) notes that in the sampling of a language,

1. the full range of registers in the language should be included, representing the range of situational variation
2. a representative sampling of texts from each register should be included; and
3. a wide range of linguistic features should be analysed in each text, representing multiple underlying parameters or variation.

Here Biber argues for the representation in a corpus of the intricate varieties of a language under study, for if a corpus lacks the major text types, genres or dialectal varieties, it cannot be said to represent the general language. Furthermore, Leech argues that:

> The value of a corpus as a research tool cannot be measured of brute size. The **diversity** of the corpus, in terms of the variety of registers or text types it represents, can be an equally important (or even more important) criterion. So, too, can the care with which it has been compiled…" (Leech, 1997: 2, emphasis in the original).

Register diversity is therefore crucial in a corpus to ensure the faithful representation of linguistic variability found in a language.

While Biber et al. argue against proportional representation, Rayson (2002: 42) contends that for a corpus to be representative of the language as a whole, it should contain samples of all major text types and, "if possible, be in some way proportional to their usage in every day language." This sense of representativeness is different to that suggested by Biber (1994 and 1998) since while he argues for the inclusion of the diversity of text types in a corpus; Rayson argues that such samples should be in some way proportional to the varieties used in a language.

Corpus linguists and corpus lexicographers consistently argue for representativeness in corpus construction mainly because for corpus results to be generalized to the whole language, the corpus must be seen to be compiled in a systematic manner that is perceived to be representative of the population from which it was abstracted to justify the generalizations. Summers points to the functionality of corpus representativeness when she says:

> One of the many reasons for wanting the corpus to be representative was so that reliable frequency statistics could be generated and used to aid the lexicographers in making the many linguistic judgements that lie behind the final entry for a word in the printed dictionary (Summers, 1996: 261).

The lexicographer's linguistic judgements aided by frequency statistics that Summers refers to, include amongst other things how to frame an entry, the ordering of definitions in the entries and the sub-entries of a headword (see Chapter 3, section 3.5). Such authoritative decisions may be reached through the exploitation of corpora.

Biber also expresses a similar position to that of Summers. He argues that "a corpus must be representative in order to be appropriately used as the basis for generalisations concerning a language as a whole" (Biber, 1993: 243).

It is clear that a representative and balanced corpus must represent the different genres of language use in a language community. According to those who argue for proportional sampling, a representative and balanced corpus would additionally attempt to capture the proportions, that is, different ratios of the different varieties in a specified language community. The determination of proportions is hard to achieve, as Biber (1998) has shown mainly because it is difficult to know precisely all the text types and their proportions of use in a population with its ever-changing dimensions. The difficulties are compounded when one faces the compilation of a corpus of spoken language. This is the case since as Kilgarriff (1997: 137) points out dialectal varieties stand at different ratios to one another and should be represented within a corpus that attempts to accurately capture the language dimensions as a whole.

4.2.2 A cautious approach to balance and representativeness

On the other hand, Kennedy is not convinced that the representativeness ideal can be achieved in a corpus.

> The extent to which a corpus can ever be considered to represent a language in general is currently a matter of some contention. In practice, whether a finite sample of a language could ever 'represent' the vast amount of a language produced in even a single day is always likely to be, in the final analysis an act of faith (Kennedy, 1998: 21).

Kennedy (ibid: 62) is additionally doubtful that we can confidently argue for representativeness of a corpus that represents a language.

> In light of the perspectives on variation offered by several decades of research in discourse analysis and sociolinguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even

> of a particular genre or subject field or topic (Kennedy, 1998: 62).

By "perspectives on variation" Kennedy refers to different speech varieties that exist in a speech community. He is referring to challenges faced by sampling the standard against non-standard varieties; various sociolects covering socioeconomic status, gender, ethnicity, age, occupation, and others; different regional varieties, like *Sengwaketse*, *Sekgatla*, *Sekwena*, *Sengwato* in the case of the Setswana language; different registers like casual, formal technical and others. Such variations are difficult to represent in a corpus. By noting this difficulty, Kennedy does not imply that representativeness should not be attempted, but that perhaps theoretically an attempt at representativeness may not conclusively capture the nuances of existing varieties as perceived in linguistic research. He therefore concludes that "a 'representative' sample is at best a rough approximation to representativeness, given the vast universe of discourse" (Kennedy, 1998: 52).

Rundell also reveals the practical challenges of achieving representativeness and balance:

> In practice, it is not always feasible to assemble precisely the corpus one ideally wants: practical constraints, such as a shortage of time and money, the variable availability of machine-readable text, and problems with copyright clearance, all conspire to make compromises necessary (Rundell, 1996, online[7]).

It is precisely the problems outlined by Rundell, which stand out as some of the major impediments particularly in the African context to corpus construction. The lack of machine readable data, the unavailability of funding, the demanding transcription of spoken language and cleaning of scanned texts remain as hurdles to building corpora that capture linguistic variability of a specific linguistic community.

In compiling the BNC, Burnard notes that the objective was to define a stratified sample according to stated criteria, so that while no-one could reasonably claim that the corpus

[7] http://www.ruf.rice.edu/~barlow/futcrp.html

was statistically representative of the whole language in terms of either production or reception, at least the corpus would represent the degree of variability known to exist along certain specific dimensions, such as mode of production (speech or writing); medium (book, newspaper etc.); domain (imaginative)… (Burnard, 2002: 60).

Burnard emphasises the difficulty of attempting linguistic representativeness in a tight statistical sense, but rather that corpus representativeness for the BNC was determined in terms of known linguistic varieties, a position similar to the one held by Biber (1994).

A corpus intended to represent the "general language" but lacking in linguistic variability can lead to erroneous conclusions. Ooi argues that "a corpus selected wrongly or inadequately runs the risk of generating not only 'noise' in the information acquired but not offering any information at all" (Ooi, 1998: 52). Take for instance Verlinde and Selva (2001) who compare the corpus-based and intuition-based lexicography in French lexicography. They note that although the French lexicographers were some of the first to incorporate corpus approaches to dictionary making the lexicographic landscape in France has largely remained intuition-based. They use 50 million words of the 1998 issues of *Le Monde* and *Le Soir* to draw up a frequency list and make comparisons between the corpus list and dictionary entries. For their electronic French learner's dictionary they decided to limit the selection of their lemmas to 12 156 words by including only those lemmas that occurred at least 100 times in a 50 million-word corpus. Since it is a learners' dictionary certain words were excluded. Amongst these were words found in current affairs like *bosnique*, *kosovar* and *brainois*. By running frequency lists they identified that 12% of the 12,000 most frequent words of their corpus did not occur in *Dictionnaire du français*. They thus concluded:

> Corpus-based lexicography gives strong and necessary empirical evidence to the lexicographer's personal intuition, even if this personal intuition remains helpful in filling the gaps in our corpus (Verlinde and Selva, 2001: 598).

While they make a valid point concerning corpus-based lexicography, at least one point of criticism may be made in relation to Verlinde and Selva's experiment on the basis of the nature of the corpus they used.

Although they admit that central to corpus building are the matters of corpus representativeness and size, for them to "rely on the texts that are freely accessible" (Verlinde and Selva, 2001: 594) and in this case, text from two newspapers, defeats the point of representativeness that they attempt to defend. Biber arguing for his Multi-Dimensional (MD) approach to studying language variation has shown that a single register cannot be said to represent broad linguistic variability of a language.

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

Biber's view equally applies to corpora designed for lexicography. An admission with qualification by Verlinde and Selva (2001: 594) that: "We cannot say that our corpus is perfectly balanced, but it is made up of the kind of texts that the potential users of our dictionary will have to deal with" undermines the linguistic variability found in different genre and text types since the 50 million-word corpus is highly skewed towards one kind of genre, namely, newspaper text. Their frequency lists are not compelling although extracted from a huge corpus. The corpus lacks texts from domains such as novels, magazines, radio interviews, textbooks, sports commentaries, film, poetry, speeches and spoken text, which we expect dictionary users to encounter daily. Since their corpus lacks text variability, their criticism of *Dictionnaire du français* that it lacks certain words found in their frequency list may only be because of the inadequacy of their corpus rather than the introspective lexical inclusion principle on the part of *Dictionnaire du français* compilers. Verlinde and Selva could have evaluated their list to ascertain that it captured words from cross the spectrum of French language use. Additionally, research needs to be conducted on the degree of linguistic variability in newspaper text compared to corpora compiled from a variety of text types.

Sinclair (2004) cautions against claims of mathematical exactness in language sampling by arguing that,

> We should avoid claims of scientific coverage of a population, of arithmetically reliable sampling, of methods that guarantee a representative corpus. The art or science of corpus building is just not at that stage yet, and young researchers are being encouraged to ask questions of corpora which are much too sophisticated for the data to support. It is better to be approximately right, than to be precisely wrong (Sinclair, 2004).

Sinclair's position does not mean that he opposes representative corpora or that corpora cannot be representative, for he argues that "The contents of the corpus should be chosen to support the purpose, and therefore in some sense represent the language from which they are chosen" (Sinclair, 2004). However what he opposes is the assumption that the population is well defined, fully known and perfectly understood.

Sinclair (2004) also argues that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within and outside it that cause it to develop continuously. As a consequence he contends that no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. This position is similar to that of Biber et al. and Kennedy discussed earlier who argue respectively:

> ….we do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all variation in texts (Biber et al., 1998: 246).
>
> and
>
> …. it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic (Kennedy, 1998: 62).

Sinclair therefore argues that corpora researchers sample, like all the other scholars who study unlimitable phenomena. He argues that:

> We remain, as they (scholars who study unlimitable phenomena) do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population (Sinclair, 2004: online[8]).

By arguing against proportional representation Sinclair agrees with Biber et al. (1994) who argue for stratified and non-proportional sampling.

He argues that to discuss the concept of representativeness we must consider the users of the language we wish to represent and ask ourselves the following questions:

- What sort of documents do they write and read, and what sort of spoken encounters do they have?
- How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications?
- How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web-pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence?
- How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

(Sinclair, 2004: online[9])

Such questions will guide a compiler in selecting relevant text to include in the corpus.

Sinclair (2004) again is helpful in pointing out that “The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.”

[8] http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

[9] http://www.ahds.ac.uk/creating/guides/linguisticcorpora/chapter1.htm

Kilgarriff and Grefenstette (2003: 340) echoing Kennedy (1998: 62) argue that "representativeness" begs the question "representative of what?" The problem of what is represented by corpora is particularly compounded by designs of corpora of "general language" which is hard to define. Representativeness therefore raises serious theoretical issues about language modelling including issues such as:

- *Production and reception*: is what is modelled received (read and heard) or produced (written and spoken) language or both? The British National Corpus, for instance, attempted to take care of both perspectives (Burnard, 2002: 22).
- *Balance between speech and text corpus amounts*: We must also contend with whether spoken text can be accurately sampled and represented along the same lines as written text. How many words are we looking for and what percentage of the spoken language do such words constitute? Whether spoken text can be sampled in any representative manner is greatly questionable. While we can sample *Sengwaketse*, *Selete, Sengwato, Sekwena,* or *Sekgatla* dialects in the Setswana language, establishing an acceptable representative percentage of the spoken form of these dialects poses great difficulties since as we attempt to quantify them, more speech instances are produced. Even if we settled for a stratified sampling, we are left with the question of, how much from each stratum?
- *What constitutes distinct language events*? Do repetitions, copying, quotation, or republications of similar stories in different newspapers constitute distinct language events that could be represented in a corpus?

With the haze that clouds matters of representativeness and balance, and with limited understanding of text types, genres language varieties in research, Kilgarriff and Grefenstette, writing about using Web text as corpus, argue that:

> The web is not representative of anything else. But nor are other corpora, in any well-understood sense. Picking away at the question exposes how primitive our understanding of the topic is, and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it may be modelled (Kilgarriff and Grefenstette, 2003: 343).

Kilgarriff and Grefenstette argue that corpora if well understood cannot be said to be representative of anything else.

So far we have attempted to show the complexity of matters of balance and representativeness and how researchers differ on whether language can be sampled in a represented manner. As Sinclair (2004) has noted, one major complicating factor in building balanced and representative corpora is that language is an "unlimitable phenomena". It is unknown how many words or sentences exist in writing or how many have been uttered or will be uttered. A quest to quantify such data would result in general estimates, for more publications are produced every minute and speech is continuously produced. Such recognition of language as an unlimitable phenomenon however does not obstruct researchers from arguing for sampling different linguistic varieties for both quantitatively and qualitatively inspection. The challenge for corpus linguists and lexicographers is to identify the parameters of a language to be studied and sample them for corpus analysis. Sinclair (2004) suggests the following ways of achieving representativeness in a corpus:

1. decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components;
2. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;
3. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type;
4. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;
5. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;
6. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

(Sinclair, 2004: online[10])

[10] http://www.ahds.ac.uk/creating/guides/linguisticcorpora/chapter1.htm

While it may be difficult to define and accurately characterise balance and representativeness, most modern corpus based lexicography research still consider issues of representation and balance (Ooi, 1998) as marks of standards of authenticity and robustness in corpus construction as Sinclair shows:

> The notion of balance is even more vague than representativeness, but the word is frequently used, and clearly for many people it is meaningful and useful. Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements (Sinclair, 2004).

Reményi (2001: 486) argues that "the problems of 'representativeness' are mostly due to the double nature of the unit of observation in corpus design: either the diversity of language users, or that of text types is eclipsed." The problems lie in whether language users (text producers and receivers) or texts (the products of language use) be chosen as the units of observation. Additionally corpora organised by demographic proportions would not support the criterion of 'sample variability matching population variability' as far as text types are concerned.

Atkins et al., introduces the concept of *organic corpora*, as a possible approach of addressing matters of representativeness and balance.

> A corpus builder should first attempt to create a representative corpus. Then this corpus should be used and analysed and its strengths and weaknesses identified and reported. In the light of experience and feedback the corpus is enhanced by the addition or deletion of material and the circle repeated continually. This is the way to approach a balanced corpus. One should not try to make a comprehensive and watertight listing […] rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language […] In our ten years' experience of analysing corpus material for lexicographic purposes, we have found any corpus – however unbalanced – to be a source of information and indeed inspiration. *Knowing that your corpus is unbalanced is what counts* (Atkins et al., 1992: 10, italics mine).

Atkins et al.'s approach is attractive since it recognizes language as a growing and living entity which must be equally matched with a vibrant and growing corpus. Their position is shared by Čermák, who argues that,

> Thus it is hard to see why most (almost all) corpora are seen as strictly time-limited projects only which, when finished and having served their purpose, are far from being maintained, modernized, and substantially enlarged.... Since any language needs a consistent, perpetual, and next-to-exhaustive coverage of its data, it should have a corpus of corresponding qualities… This is particularly important in the case of minor languages which, unlike English and other languages, cannot afford the luxury of having a variety and multitude of corpora, at least not at the moment (Čermák, 1997: 182).

However both Atkins et al. and Čermák do not claim to have solved the matter of balance, rather they argue for a constant updating of the corpus over time – a position similar to that of Sinclair (1989: 29) who points out that "…a corpus should be as large as possible and should keep on growing". Even if a corpus is updated continuously, the challenge will remain in that some corpus linguists would want to work with a finite and constant entity such as the BNC rather than an entity whose contents are in perpetual flux.

It should be fairly clear that what constitutes balanced and representative corpora still remains controversial. The matter of how much sampling of different genres to include in a corpus is still largely unresolved. "The crux of the matter is finding a criterion for selecting the proportions between the reception and production" of text (Čermák, 1997: 192). What appears to be agreed upon though is that a corpus must finally capture the language varieties from a specified population from which a sample is taken, which reflects how that particular language community uses language. This is significant since (Summers, 1993: 186, 190) argues that the results of corpora analysis may be generalised to the general language community from which the samples were abstracted and Kennedy (1998: 94) shows the results of corpus analysis may have pedagogical function since "high frequency of occurrence as determined by the analysis of texts should be a major determinant of lexical content of language instruction".

Issues surrounding the exploration of linguistic variability have engaged many other researchers (Kittredge, 1982; Zwicky and Zwicky, 1982). Since corpora that substantially cover the full range of registers have been shown to be invaluable to both lexicographic research and studies in language variation, we are compelled that the corpus models for the Setswana language and other languages ought to represent a range of register diversity in both spoken and written situations.

4.3 Corpus annotation

Having collected texts into a corpus, such a corpus can contain simple raw text or it can be enriched with linguistic information before information extraction. The raw text can also be annotated or marked up. The mark-up language is concerned with the encoding of a corpus. The encoding, referred to as annotation or tagging, added to the texts that comprise a corpus, is a metalanguage that is generally done in some form of mark-up language (Horvath, 1999: Section 2.3.1). Two commonly used mark-up languages in corpora are XML and SGML. The Extensible Mark-up Language (XML) is the universal format for presenting structured documents and data on the World Wide Web (WWW). The functionality of the Web is improved through XML's design because it provides more flexible and adaptable information identification. "It is called extensible because it is not a fixed format like HTML (hyper-text mark-up language), which is a single, pre-defined mark-up language" (Pravec, 2002: 101). As a metalanguage, XML allows the design of customized mark-up languages for a limitless number of different types of documents. This is made possible because it is written in Standard Generalized Mark-up Language (SGML), the international standard metalanguage for defining descriptions of the structure for different types of electronic documents.

Grammatical tagging is one common practice of adding interpretative linguistic information to a corpus at various levels (Monachini and Picchi, 1992). It classifies each word-form in a text, labelling it with a part of speech tag (POS-tag) and morphological features. The process can be performed automatically. The part of speech mark-up is particularly crucial. De Rose (1991: 9) has shown that 11% of word types and 48% of word tokens occur with more than one category label (Kennedy, 1998: 209). For instance, the mark-up of the sentence: "There is nothing masculine about these new trouser suits in summer's soft pastels." from the BNC (Burnard, 1995: 35) follows below:

```
<s n=00041>
<w EXO>There <w VBZ>is <w PNI>nothing <w AJO>masculine
<w PRP>about <w DTO>these <w AJO>new <w NN1>trouser
<w NN2-VVZ>suits <w PRP>in <w NN1>summer<w POS>'s
<w AJO>soft <w NN2>pastels<c PUN>.
```

The POS-tags in the above sentence are to be understood as follows:

AJO	: Adjective
DTO	: general determiner
EXO	: existential there
NN1	: singular common noun
NN2	: plural common noun
PNI	: indefinite pronoun
PRP	: preposition, other than *of*
POS	: the possessive or genitive marker *'s* or '.
VVZ	: the –s form of lexical verbs, e.g. *forgets, sends, lives, returns*
PUN	: any mark of separation (.!,:;-?..)
<s>	: segment
<w>	: word
<c>	: a punctuation mark

The part of speech annotation can also be parsed or marked for syntactic information to show the phrase, clause or sentence divisions. The Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1994) is a sophisticated attempt at establishing guidelines of how to encode machine-readable text through a complex application of SGML. The SGML was used in the mark-up of the BNC which uses the Corpus Development Interchange Format (CDIF). This international standard provides, amongst other things, a method of specifying an application-independent document grammar, in terms of the elements which may appear in a document, their attributes, and the ways in which they may legally be combined (Burnard, 1995: 25). The detail of the mark-up is only relevant to the function to which the corpus would be put to as Kennedy (1998: 84) shows: "The level of detail of mark-up has to be related to the potential use of the corpus." Programs such as CLAWS (Constituent Likelihood Automatic Word-tagging System) (Garside and Smith, 1997) have also been used in tagging various corpora like the BNC (see BNC website[11]).

.

Tagged corpora are useful in corpus linguistic research in that they can help in the development of disambiguation rules and facilitate automatic and semi-automatic syntactic analysis. Tagged corpora have also been found to be highly useful in the generation of word sketches. "Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004: 105).

Kilgarriff and Rundell (2001: 807) show that as corpora grow, so does the number of corpus lines for a word. This leads to what they call "the problem of information overload" for a lexicographer when he or she has to deal with a great number of concordance lines. The solution lies in statistical summaries. Kilgarriff and Rundell (2001) have generated word summaries through "Word Sketch" software which uses parsed corpus data to identify salient collocates – in separate lists – for the whole range of grammatical relations in which a given word participates (see also Kilgarriff and Tugwell, 2000). They report that lexicographers found that the Word Sketches not only streamlined the process of searching for significant word combinations, but often provided a more revealing, and more efficient, way of uncovering the key features of a word's behaviour than the method of scanning concordance lines. They offer detailed information that would be hard to extract from a corpus which is not annotated. We illustrate this with the word sketch for *pray* from Kilgarriff et al., (2004: 120).

Figure 2: Word sketch for pray (v)

pray (**v**) BNC freq= 2455

miracle	8	13.9	emperor	2	5.2	read	9	9.5	inwardly	3	5.5	hook	2	3.3	she	130	5.8
for him	**680** 26	**3.4** 13.7	**to** Jesus	**142** 2	**4.5** 4.1	**and/or** talk	**179** 6	**1.7** 7.4	**modifier** hard	**338** 7	**0.5** 9.5	**object** thing	**183** 3	**-1.2** 3.2	**subject** muslim	**1361** 3	**0.7** 9.5
forgiveness fair	12 7	19.8 19.4	spirit god	32 2	24.0 4.3	hope sing	20 4	20.8 6.4	silently daily	15 3	13.3 4.4	god night	53 3	10.5 3.1	follower eve	306 3	12.3 5.0
soul you	14 23	19.3 19.2	God image	22 2	17.0 4.7	hope watch	43 4	15.5 5.0	together only	20 35	9.8 9.3	God lord	21 2	9.7 2.6	petitioner jesus	3 7	8.8 4.3
me -	147 24	17.1 13.3	lord wind	16 2	11.4 3.9	fast live	6 3	12.2 3.9	fervently continually	4 3	7.6 3.7	prayer pardon	6 2	7.6 2.7	knee jew	5 3	6.9 4.5
God deliverance	61 6	16.0 15.5	saint him	4 6	10.0 3.3	pray work	16 5	11.2 3.5	aloud regularly	6 5	7.5 3.5	day soul	9 2	3.8 2.4	congregation church	4 7	6.8 4.3
peace then	25 23	16.5 12.2	jesus	2	5.4	kneel wish	5 2	9.9 3.4	earnestly often	10 5	7.3 3.3	heaven silence	3 2	3.4 2.2	fellowship I	263 2	6.0 4.2

[11] http://info.ox.ac.uk/bnc

church	12	11.7	believe	2	2.9	ever	9	3.0	Singh	2	3.7
guidance	8	11.6	learn	2	2.8	secretly	2	2.7	Family	6	3.6
us	16	11.6	tell	2	2.3	quietly	3	2.4			
chance	5	10.3				still	11	2.3			

The Word Sketch therefore helps reveal that people usually pray for *rain, soul, God, peace, peace miracles, forgiveness* amongst other things. It also reveals that the verb *pray* is usually modified by *silently, together, fervently, aloud* and *earnestly*. Such wealth of information would have been difficult to uncover without the help of Word Sketches.

4.4 Sample size

Every corpus is a language sample (Leitner, 1992). As discussed earlier (Chapter 3) a corpus can comprise sampled text from books, newspaper, speech and other text. Other corpora comprise complete works of writers, or complete texts such as the Bible, but they also in a sense constitute samples of language use by such writers or of particular genres. Such corpora will be discussed briefly later. What must be established foremost is that text sampling is central and basic to corpus construction. This position finds support in Biber, who points out that,

> Some of the first considerations in constructing a corpus concern the overall design: for example the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples within texts and the length of text samples. Each of these involves a sampling decision, either conscious or not (Biber, 1994: 377).

The matter of sample size is closely related to the previously discussed subject of representativeness since the number and size of texts in a corpus determine whether a corpus can be judged as representativeness of a language or not.

The purpose of sampling adequately is so that reliable generalizations may be made concerning a population as a whole. However, as we have seen, a linguistic population is normally so large (in terms of the number of speech acts produced) and so indefinable (in terms of the possible range of text types) that a random sample, stratified according to all major language text types, is probably not feasible (Kennedy, 1998: 74).

In corpus compilation one issue that still needs to be explored is how much of each text type sample should be included in a corpus. For those compiling opportunistic corpora, any amount of text found may be added to the corpus. For those attempting balanced corpora the need to define the population to sample becomes urgent and a decision of how much text from each text type must be made. However the language to be sampled, such as Setswana, as Clear (1992: 21) has argued, is poorly defined. Unlike in other studies where the population is clearly defined, say university students or people over the age of fifty, something like the Setswana language is not perfectly defined. It is broad with a variety of dialects; it is not clear whether we refer to produced (books, speech, etc.) or received language (language that we hear or read). It is also not clear what unit of language is best to be sampled and analysed, that is, whether we are interested in sampling words, sentences or whole texts such as books or conversations. The challenge that arises in sampling is that there is a real possibility that one may under-represent some variety of language in a corpus as Clear has shown:

> Given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample (Clear, 1992: 21).

Although defining a population to be sampled is difficult, it however has to be done if generalisations drawn from a corpus are to be made about a broad language community.

Different corpus compilers sample language differently. The Brown Corpus and the Lancaster Oslo Bergen (LOB) corpus each has 500 samples of 2,000 words each. Sinclair (1991: 19) argues that the even sample sizes are advantageous as far as making comparisons is concerned. In the BNC case a target sample of 40,000 words was chosen for books and anything less than 40,000 was reduced by 10% for copyright reasons (Burnard, 1995: 10).

Sinclair (1991: 19) points out that an alternative to smaller text samples is “to gather whole documents” and adopt a policy of continuous corpus growth since “from a large corpus can be drawn any number of smaller, more specialized ones, according to requirements from time to time.” The weakness of collecting whole documents as a collection strategy is that the coverage will not be as good as a collection of small samples and one text characteristics may dominate others. On the size of a corpus sample, Biber (1995: 132) concludes that "1,000-word samples reliably represent many of the surface linguistic characteristics of a text, even when considerable internal variation exists."

Kennedy (1998: 20/21) argues that complete works corpus is "not representative of an entity. It is that entity."

De Haan (1992: 1) points out that one thing that has not been explored is how the size of corpus samples affects the research results. From a variety of experiments he conducts, he shows that the suitability of a sample depends on the specific study that is undertaken, and as if answering Biber's (1995: 131) question "What is the optimal text sample length?" he argues that there is no such thing as the best, or optimum, sample size.

Leech (1991: 10) argues that a preoccupation with size "…is naïve – for four reasons."

1. A collection of machine-readable text does not make a corpus. A corpus has to be designed for a specific representative function.
2. The vast growth of resources of machine-readable text has taken place exclusively in the medium of written language – speech devices have not developed the automatic input of spoken language to the level of the present OCR (optical character recognition).
3. While technology advances quickly, human institutions evolve slowly. Problems relating to copyright forbid the copying of text without the license of the copyright holder. It is therefore difficult to find corpus that is available unconditionally for all users.
4. While hardware technology advances, software technology lags behind. Having enormous amounts of text but lacking the software to explore them is unfruitful.

Leech shows that brute size in corpus compilation is not everything. The corpus must be representative; representing written as well as spoken language. He observes that developments in software technology will go far in aiding information retrieval from corpora.

The brief discussion of sample size is aimed at showing that while sampling lies at the heart of corpus compilation, different corpus linguists adopt different sampling approaches. The Brown Corpus and the LOB corpus each has 500 samples of 2000 words each. The BNC comprises samples of 40,000 words for books and anything less than 40,000 has been reduced by 10% for copyright reasons (Burnard, 1995: 10). For those compiling opportunistic corpora, any amount of text found may be added to the corpus. It appears that the purpose to which a corpus would be used for need defining prior to any sampling. If a corpus is to be used to compare equal text samples then sampling chunks with equal number of words may be a desirable option. However in NLP, an opportunistic corpus may be ideal; while for lexicography, a corpus with broad coverage is desirable (see Manning and Schütze, 1999).

4.4.1 Spoken versus written corpus text

Speech in a language community is the primary channel of human communication and exists in abundance compared to written text (Cho and O'Grady, 1996). While this is common knowledge in linguistics, language researchers do not know quantitatively how much of speech exists, nor do they have the resources and methodologies to account for how many words are spoken daily by interlocutors.

General language corpora in order to better represent a language it must include both spoken and written text, different text genres and various dialectal varieties. If a corpus is compiled proportionally then spoken language would be greater than written language in a corpus. However this does not hold true in many corpora compilations since some are not sampled proportionally but in a stratified manner (see Section 4.2.1). Sinclair (2004) points out that "estimates of the optimal proportion of spoken language range from 50% — the neutral option — to 90%, following a guess that most people experience many times as much speech as writing." Such a greater occurrence of spoken text over the written one would approximate the ratios of written and spoken text in the real world and would be likely to produce corpora that closely represent language as used in speech communities. However in none of the large corpora like the BNC and the

Bank of English does the percentage of the spoken text exceed that of written text. The BNC, a 100 million words corpus of modern spoken and written English, has 90% written text and 10% spoken language. The ratios between the spoken and written corpus do not approximate the real world ratios of linguistic differences between spoken and written language. Sinclair (2004) argues that "most general corpora of today are badly balanced because they do not have nearly enough spoken language in them." This is true of the BNC although the BNC is one of the corpora with the largest spoken text (about 10 million words). Such an imbalance raises questions relating to the composition and balance of the corpus and also points to the fact observed by Sinclair (2004) that a corpus is an imperfect entity. He argues against any exactness in corpus compilation thus:

> 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).

Leech et al. also recognise the inadequate representation of speech in the BNC thus:

> Although spoken language, as the primary channel of communication, should by rights be given more prominence than this, in practice this has not been possible, since it is a skilled and very time-consuming task to transcribe speech into the computer readable orthographic text that can be processed to extract linguistic information. In view of this problem, these proportions were chosen as realistic targets which, given the size of the BNC, are also sufficiently large to be broadly representative (Leech et al., 2001: 1).

According to Leech et al. the percentage of speech text in the BNC, was reached by determining what was possible to the compilers and not as a consequence of proportions of speech to written text in the English language. BNC designers could have arrived at the 90% and 10% ratios by studying the language situation of a speech community and projecting the estimated ratios of spoken and written language into the corpus structure. But according to Leech et al. these ratios were purely 'chosen as realistic targets' of limitations in the spoken language transcription and because of the expensive nature of manual transcription.

It is not clear if a situation in which a corpus has more spoken language is desirable for linguistic analysis. Biber (1994) has argued that to have greater spoken language percentages in a corpus is not linguistically interesting since the corpus ends up being homogeneous. What corpus compilers should aim for, he argues, are stratified corpora that capture the linguistic variability of the language community and not proportionally-compiled corpora. This position has however been rejected by Varadi (2001) who prefers proportional sampling and accuses Biber of attempting to redefine representativeness by divesting

> …such a key term of its well-established meaning, which has a clear interpretation to statisticians and the general public alike… There is such a strong and unanimous expectation from the public and scholars alike for corpora to be representative that it is an assumption that is virtually taken for granted. However, to meet this demand by the semantic exercise of redefining the content of the term is a move that hardly does credit to the field (Varadi, 2001: 592).

There are added challenges to spoken corpus compilation. It is not only the matter of what it means to be representative as seen in the different position taken by Biber and Varadi above. Atkins et al. express frustrations with building a corpus of spoken text when they say:

> The difficulty and high cost of recording and transcribing natural speech events lead the corpus linguist to adopt a more open strategy in collecting spoken language (Atkins et al., 1991: 3).

Atkins et al. suggest that technological inadequacies in speech transcription force corpus linguists to settle for corpora that are not desirable, but tolerable. Such positions inevitably raise the theoretical questions of whether corpus representativeness could be sustained in conditions in which the desired and representative corpora do not exist (see also Rundell, 1996).

Sharoff proposes one way of solving the lack of spoken material in a corpus in this way;

> The proposed solution is to increase the amount of ephemera (including leaflets,

> junk mail and typed material), correspondence (business and private) and spoken language samples whenever possible, because they reflect everyday language produced and reproduced regularly in discourse (Sharoff, 2004: 6).

Sharoff's attempts at solving the impasse illustrate the gravity of challenges of compiling spoken language. However the extent to which material such as business and private correspondence, leaflets and junk mail can substitute for spoken language is still to be investigated.

4.4.2 Newspaper text versus the purchase of a pair of shoes

Other researchers look at the matter of spoken language representation in a corpus differently. While they acknowledge the common occurrence of speech in daily discourse, they argue for more written language in a corpus since they consider speech private and restricted to a few interlocutors, while written text such as novels and newspapers have broad readership and deserve prominence in a corpus. One researcher who holds this view is Kennedy who argues:

> No one knows what proportion of the words produced in a language on any given day are spoken or written. Individually speech makes up a greater proportion than does writing of the language most of us receive or produce on a typical day. However, a written text (say in a newspaper article) may be read by 10 million people, whereas a spoken dialogue involving the purchase of a pair of shoes may never be heard by any person other than the two original interlocutors (Kennedy, 1998: 63).

Kennedy introduces an interesting dimension to corpus compilation that raises great controversy. It is true that a newspaper is likely to be read by many people and that its circulation may be verified from reliable sources. However the challenge still remains since newspaper buyers do not read the same sections of a newspaper. Some people have no time for business section, classified, cartoons, letters to the editor and other newspaper sections. Although circulation numbers might be available to assist corpus builders sample newspaper text, they only give numbers of purchased newspaper but do not quantify patterns of newspapers readership.

A similar point may be made that although many of the corpora depend on published texts, there is indeed no guarantee that such texts are widely read (or read at all). This is particularly so in the Setswana language situation where the majority of Batswana do not read Setswana text, save in Setswana classes at both primary and secondary school. Kennedy (1998: 52) suggests that to fix this problem "best seller lists, library lending, statistics and periodical circulation figures can only partially reflect receptive use and influence." Kennedy's use of *partially* is an indication of the immensity of problems surrounding attempts to construct corpora on the basis of common and influential text. If "receptive use and influence" are taken as determinants of text inclusion in a corpus we must contend with varying *degrees* of such use and influence. School textbooks and creative texts read by thousands of students would be in use more than a library text that is rarely read. It is not clear how such a distinction will be reflected in corpus compilation. Textbooks would have been read more widely and therefore their text should somehow reflect the fact that they have been seen more than other texts. The argument may be pushed further. This would mean that a sign that reads: "Welcome to Gaborone" would make "welcome" "to" and "Gaborone" more common since such language would have been received by many people. However it is not clear how such information could be represented in a corpus. What about words like "Stop" used in traffic signs and seen by people repeatedly daily? Arguing for more of written language in a corpus since written language has been read widely or seen repeatedly compared to spoken language which is private, makes the discussion complex and in no way resolves challenges of the representation of speech in a corpus.

It would appear that Kennedy's argument against spoken text on the basis that it is private while written text is in the public domain, is not convincing but rather raises new problems and challenges as outlined above. Spoken text is as important as written text in corpus compilation and novel attempts need to be made to achieve its better representation in a corpus.

4.4.3 The value of spoken language

In this section we illustrate the value of spoken language to corpus research. We illustrate what might be missed if a corpus does not include spoken language text. Borrowings and colloquialisms are common in speech but they are dispreferred by editors and publishers, especially in communities where there is language contact such as the African context. Spoken Setswana is characterised by high levels of borrowing from English and Afrikaans. The documentation of foreign acquisitions in Setswana is not recent. Cole (1955) noted words like *beke* from “week”, *baki* “*baadjie*” (jacket), *gouta* from “*goud*” (gold), *heke* from “hek” (gate), *hempe* from “*hemp*” (shirt), *kofi* from “*koffie*” (coffee), *pena/e* from “pen”, *peipe* from “*pyp*” (pipe), *sukiri*, from “*suiker*” (sugar) and *baesekele* from *“*bicycle*”, buka* from *“book”, ofisi* from “office”*, šeleng* from “shilling”. There are other more recent borrowings like *gate* which reveal a certain layering in the nature of borrowed words. For instance, many Setswana speakers do not recognise *jase* from “jas” (coat)*, heke* (hek) “gate” and *baki* (baadjie) “jacket”, as borrowings from Afrikaans, while *jakete* (jacket) is recognised as borrowed from English. There is a similar situation with *heke*, which is considered by some speakers of Setswana as a sign of ‘good old Setswana’ while *geiti*[12] is recognised as an obvious borrowing. Spoken Setswana is peppered with instances of borrowing, code-switching and colloquialisms as illustrated in the following sentences.

Spoken Language	**English Equivalent**
Go shapo!	Bye
O tsile ka thelebišene	He came with a television
Ke bra/sistere ya gagwe.	It is his brother/sister.
O apere jase.	He is wearing a coat.

In the above examples *thelebišene* is a borrowing from the English noun *television* and *jase* from the Afrikaans noun *jas* and *shapo* a colloquialism which means *fine* or *bye*.

Borrowing and code-switching can also be seen in dialogue including days of the week, months and numerals. For instance, many Setswana speakers would say *Monday* or

[12] *geiti* is borrowed from the English "gate". Since Setswana does not have the voiced, velar plosive as part of its sound system, which in this instance occupies the initial word position in *geiti*, there is no agreed orthographic representation of such a sound in Setswana.

Mantaga (from Afrikaans, *Maandag), Saturday* or *Sateretaga* (from Afrikaans *Saterdag*) *and Sunday* or *Sontaga* (from Afrikaans, *Sondag*).

Setswana speech is also characterised by high degrees of code-switching, speakers switching from Setswana to English. This is particularly common in the use of English numerals in many instances instead of Setswana terminology. Many Setswana speakers would have difficulty in saying 1,567 in Setswana (i.e. *sekete, makgolo a matlhano le masome a marataro le bosupa*). Numbers are generally said in English. It is common for Batswana to use *one, two, three, fifteen, two thousand,* or *one million,* in their speech instead of Setswana terms *bongwe, bobedi, boraro, lesome le botlhano, dikete tse pedi* or *sedikadike*, respectively. Take the example below of a dialogue about selling. The example is from the spoken component of the Setswana corpus that we have compiled. English translations are given in brackets and numbers in Setswana speech have been italicised.

Dialogue 1

MT: Shess... A a! ka nne ke letse ke bua le ene. A bo o mo neela ka *one fifty*. (Wow! But I was speaking to her yesterday. And you gave him for one fifty.)

TP: *One fifty*?

MT: Ee (Yes)

TP: O ne a re wa re *sixteen* Pula. (She said you said sixteen Pula)

MT: ...Ke ne ke re, ka re *sixteen fifty***.** (I was saying, I am saying sixteen fifty)

TP: Ee, ke be ke mo neela ka *sixteen fifty*. Go tlhaela *six* Pula... (Yes, I then gave her for sixteen fifty. It is six Pula short.)

MT: O a tlhaela? (It is short?)

TP: Ee, a ke re ke ne ke mo tšhentšhetse ka madi ame. Ke raya gore ke tlaa tla ke mo go neela. (Yes, I gave her change using my money. I mean that I will give it to you later.)

MT: Ehee. *Ok* nna ka re ke ena a sa, a sa, a sa mo ntshang. (Oh I see. OK. I thought that it was her who had, who had, who had not given the money.)

TP: Nnyaa ao! Nnyaa. (No! No!)

From the above dialogue English numerals: *one fifty*, *sixteen*, *sixteen fifty*, and *six* are used in the middle of a dialogue in Setswana instead of Setswana terms *lekgolo le botlhano, lesome le borataro, lesome le borataro le metso e e masome a matlhano* and *borataro* respectively.

It is not only English numerals which Setswana speakers usually switch to in speech, reference to months is also usually in English, and many speakers would have difficulties in stating months in Setswana. We return to this discussion later in this chapter.

Below we give two dialogues one a radio call-in program and the other from an interview television programme. The first two dialogues are from the Radio Botswana call-in program *A re bueng* (Let us talk) which is conducted largely in Setswana. The subject for the day was how certain youths abuse their parents by making difficult requests and demands, and if their demands were not met the youth threatened to commit suicide. We sample only a small part of the whole program. English words in the middle of Setswana dialogue are italicised and translations are in brackets.

Dialogue 2

RBP: *Ok*, ba bangwe, o ise o tsamaye Mogotsi, ba re thupa ke yone (Ok, others, before you go Mogotsi, say whipping is the answer).

Cal: *That doesn't solve anything and* mo go dira *to the worst* fa o…, ka na nna ke tle ke re le mo loratong a re e beye, fa o ratana le motho o bo a go raya a re: "Ke a go tlogela" O bo o re ke go rekela *something*… (That doesn't solve anything, it makes matters worse and if you can…, I sometimes say that in love relationships let us put it aside, if you are in a relationship with someone and they say to you: "I am leaving you" And then you say I am buying you something…)

RBP: Mh.

Whole English sentences such as "[t]hat doesn't solve anything", phrases such as "to the worst" and words such as "ok", "and" and "something" are examples of the extent of English usage found in urban and educated Setswana speech. Below we give another speech chunk from the same call-in program.

Dialogue 3

RBP: Fa gongwe e tlaa re a tsamaile a boe. (Sometimes after he leaves, he comes back.)

Cal: A boe, mo ga go kgetla thupa o re ke betsa ngwana gore ga a batle go nkutlwa, *you are making things worse* go feta fa di leng teng. (He may come back, getting a stick to beat a child because he does not listen to me, you are making things worse beyond what they are).

RBP: Nnya mme… (No but…)

Cal: Thupa gotlhelele ga e yo tota le ko sekolong. *I don't encourage,* gore ba re thupa e ka sokolola ngwana. *Sit down* le motho, buang le ene o tlaa ipaakanya. Fa go pala go raya gore go a pala. (Whipping completely is not there at school. I don't encourage, that they say that whipping can transform a child. Sit down with someone, speak to them, they will fix themselves. If it fails, it would have failed).

Similar to the previous speech chunk investigated above, English sentences creep into Setswana speech. For instance: *You are making things worse* and *Sit down*. There are also clauses such as *I don't encourage*. We need to keep in mind that radio call-in programmes are informal programs where callers freely express their views on a variety of issues. We will however see that even in formal programmes a similar pattern of switching to the English language persists.

We now look at a formal television programme broadcast in the Setswana language. While participants in this programme come prepared to address a specific subject, they do not know the questions in advance.

The following dialogue was transcribed from the Botswana television programme, *The Eye*, which is an interactive programme with two to three interviewees tackling a current matter of concern. The subject of the program was on the drying Gaborone dam which supplies the capital city with water and the role of the Botswana Water Utilities and Water Affairs in advising and training users in water conservation.

Dialogue 4

OS: Mme se re tshwanetseng gore re se gakologelwe ke gore jaaka Mma SR a ne a bua kgantele ka gore metsi a mo matamong a kgadisiwa ke, ke *evaporation* go na le *elemente* e nngwe gape e e leng gore e teng ya gore, letamo jaaka o le itse le nna le ... mmu jaaka o ntse o tsena mo letamong o fokotsa *capacity*… (But what we should remember is that as Mrs. SR was saying earlier that water in the dam dried because of evaporation, there is another element at play, which is, the dam as you know has… as soil collects into the dam it decreases the capacity…)

MK: *So* re lebile (So, we are looking at) *eight months as the best case scenario, worst case scenario*?

GS: *Worst case scenario* mma tota re ka nna ra re (Worst case scenario, we can say) *between six and eight months.*

Dialogue 4 shows a formal educated dialogue characterised by words such as *evaporation*, *element* and *capacity* and phrases such as *worst case scenario* and *between six and eight months*. English is pervasive in spoken Setswana as Bagwasi (2003) has shown.

There are also cases of colloquialism in spoken language. An example of colloquial speech from the Setswana corpus follows (English words are bolded and colloquialisms italicised):

> Hey monna Bobi, o seka wa dira *daidee*. *Magents* bane ba tseela Tshege dilwana *daa*, a *vaela dladleng* a le maponapona, **fortunately** bane ba sa nne kgakala *plus* **it was at night.** Hey phikwe, re *chitse* ha posong *baba* gongwe ko statung (statue) rena le Comfort a nwa coca cola, **Saturday afternoon**, re planela maitseboa. Re bo re *shapa round* mo *mmolong*, re o *covera* **in 10 minutes**.O *vaa* ka **line** ya Elegant, *ga* otla o tswa ka ko Pep kakwa otla ka **line** ya Pioneer town e fedile, heish *Zana baba*.

> Hey man Bobi, don't do that. Guys stole Tshege's clothes at that place and he went home naked, fortunately they did not live far and it was at night. I remember Phikwe, we relaxing next to the post office or the statue together with Comfort drinking a Coca Cola on Saturday afternoon planning for the evening. Then we would go around the mall and cover it in 10mins. You would go from the Elegant side coming from the side of Pep stores, the side of Pioneer and you would have covered the entire mall. How I miss Phikwe! (*translation mine*).

In the above quoted text *baba* (man, sir), *shapa round* (leave and return quickly), *mmolo* (mall), *covera* (cover), *vaa* (go), *daidee* (that thing), *magents* (guys), *chitse* (chilling, relaxing), *vaela* (go towards), *daa* (there), *dladleng* (home) are all colloquial Setswana words which are not used in formal texts. It is in analyzing spoken language that the colloquialisms are encountered. The presence of colloquialisms in speech lends additional support to the inclusion of transcribed spoken language in a corpus.

What we have attempted to show so far with the different dialogues and an example of colloquialisms is that the entity called Setswana spoken language is not a uniform, clean and homogeneous phenomenon. Rather it is characterised by foreignisms and colloquialisms. Borrowing, colloquialisms and code-switching are therefore some of the issues which confront Setswana lexicographers who use a Setswana spoken corpus or a corpus comprising portions of spoken data. Such lexicographers would grapple with issues relating to spoken text amongst these being:

1. The transcription of the language. Apart from it being a time-consuming process, there are tough decisions to be made on what is borrowing and what is merely code-switching.
2. If the corpus is annotated, there will be decision on what to mark-up (coughs, sneezes, passing traffic, hesitations, etc).
3. At a practical lexicographic level some of the issues that arise from including transcribed spoken language in a corpus include decisions of the kind of borrowed words to be listed in the dictionary and the kind of stylistic information derived from borrowed words.
4. The spelling of certain words on which there is no agreement.

5. Speech which is not thought through, characterised by hesitations, back-tracking and incomplete sentences.

The challenges of the treatment of borrowings in dictionaries that face a Setswana lexicographer mainly because of spoken text in a corpus are not unique to the language. Another language that faces a similar challenge is Toqabaqita, an Austronesian language spoken in the Solomon Islands.

The inclusion of spoken language in a corpus has relevance to the treatment of code-switching and borrowed words abstracted from such a corpus in a dictionary. In the subsequent section we discuss how lexicographers have addressed the challenges of borrowing and code-switching in the Toqabaqita language and how their approach sheds light to the treatment of borrowings and code-switching to the Setswana language.

4.4.4 The treatment of borrowings in Toqabaqita

Because of language contact many languages borrow words form others. This raises questions of whether such borrowed words qualify as belonging to the borrowing language and therefore deserving to be in its dictionaries. Lichtenberk (2003) in his report on the dictionary of Toqabaqita points out that the central point in determining the wordlist of a dictionary is the consideration of intended users of a dictionary, what he calls "audience", and expectations, that is, the kind of purpose the dictionary has to serve in the society. This view is shared by Zgusta who says decisions of what to include are determined by "fundamental decisions concerning the type of dictionary which is to be prepared" (Zgusta, 1971: 243). For instance if the dictionary intends to contribute to historical and comparative studies it may list archaic and obsolete words while the inclusion of loanwords may prove to be of interest to phonologists. But the larger part of Lichtenberk's (2003) paper is devoted to the discussion of inclusion or exclusion of loanwords in the dictionary of Toqabaqita. We discuss it in detail since there are comparisons which may be drawn between Toqabaqita and Setswana. Lichtenberk is confronted with a language situation where he has to make a decision of whether to include Pijin words in the dictionary of Toqabaqita since some of them fit the phonological and phonotactic constraints of Toqabaqita while others do not. Like

Setswana, Toqabaqita does not permit consonantal cluster or syllable final consonants and has a simple syllable structure of CV and V. This is exemplified in words like *kisini*, "kitchen" and *wasia* "wash". The principle that guides Lichtenberk in deciding what to include is:

> Pijin words used in Toqabaqita are listed provided they fit the phonological and phonotactic patterns of Toqabaqita, either because they fit them already in Pijin or because they have been accommodated to them. Words which do not fit the patterns are not listed (Lichtenberk, 2003: 395).

This principle excludes certain words that are in common use which in Lichtenberk's view are instances of code-mixing (Lichtenberk, 2003: 396) and not borrowing. These words include *qambrela* "umbrella" from Pijin *ambrela* and *grup* or *grupu* "group" from Pijin *grup*. They are not listed in the dictionary since they do not satisfy the phonotactic constraints of Toqabaqita. Similar to the Setswana situation, code-mixing in Toqabaqita is common, especially in numerals, months and the names of some of the days of the week and Lichtenberk argues:

> Considering such words to be part of Toqabaqita lexicon would amount to claiming that the phonological inventory and the phonotactic patterns of the language have undergone some major changes (Lichtenberk, 2003: 396)

Therefore Lichtenberk decides to restrict the matter of code-mixing to the front matter where the common but non-accommodated words would be listed. There are also problems concerning pairs of words which though accommodated from Pijin, have variants which do not conform to the phonotactics of Toqabaqita. In this instance the variant that does not conform to the phonotactic constraints is not listed. This is exemplified by *bereta* and *bret* "bread" where *bereta* is accommodated and *bret* is not listed since it is less common and not accommodated. The situation gets increasingly interesting when the non-accommodated variant is more common than the accommodated one as in *gavman* (that violates the phonotactic constraints of Toqabaqita and is un-accommodated) and *gafumanu* (is accommodated but it is infrequent). In such a case Lichtenberk ignores the most frequent used word *gavman*, since it violates the phonotactic constraints of the language, and instead chooses to enter the less common *gafumanu* on the principle that the non-accommodated variant though frequent, is an instance of code-mixing.

Lichtenberg develops other principles which govern what to list, and these are listed below:

1. "Words that belong in well-circumscribed and relatively small sets are not listed if some other members of the same set do not occur in an accommodated form and so are not listed" (Lichtenberk, 2003: 396). Such sets include numerals, days of the week and names of months.
2. A Pijin word that has been encountered only once is not listed even if it fits the phonological and phonotactic pattern of Toqabaqita.

The question of what has to be listed in the dictionary raises an issue of the boundaries of the lexicon of a language. And Lichtenberk divides the Toqabaqita into 3 categories: i) native Toqabaqita words ii) accommodated borrowings from Pijin, and iii) Pijin words used without being accommodated. Lichtenberk concludes that:

> Only the first two types are to be listed in the dictionary, which amounts to saying that only those words are part of Toqabaqita lexicon, while the non-accommodated words are not (Lichtenberk, 2003: 397).

And Lichtenberk gives proper criticism to his approach when he says:

> The principle, while explicit and applicable in a straight forward way, is nevertheless arbitrary. It gives priority to the phonological and phonotactic patterns of Toqabaqita over usage. Pijin words that are not accommodated are, by fiat, placed outside the circumference of the Toqabaqita lexicon, although by virtue of their usage they could be inside (Lichtenberk, 2003: 397).

Lichtenberk's criticism of his principles is accurate. His principles could lead to unacceptable results. Take for instance the principle that: "Words that belong in well-circumscribed and relatively small sets are not listed if some other members of the same set do not occur in an accommodated form and so are not listed" (Lichtenberk, 2003:

396) which include numerals, days of the week and names of months. While this principle might work well in reference to numerals and names of months in Setswana, the same cannot be said for days of the week. Let us consider the days of the week data in Setswana:

Table 16: Setswana days of the week

English	Standard/written	Kgasa (1976)	Spoken/Common
Sunday	*Tshipi*	*Lantlha (Tshipi)*	*Sontaga*
Monday	*Mosupologo*	*Labobedi*	*Mantaga*
Tuesday	*Labobedi*	*Laboraro*	*Labobedi*
Wednesday	*Laboraro*	*Labone*	*Laboraro*
Thursday	*Labone*	*Labotlhano*	*Labone*
Friday	*Labotlhano*	*Laborataro*	*Labotlhano*
Saturday	*Matlhatso*	*Labosupa (Sabata)*	*Sateretaga*

Table 16, shows days of the week in Kgasa (1976), in common spoken language and in standard written Setswana. Standard Setswana names are used in text books, novels, and government media and in creative writing in schools. In the table the column with standard Setswana is followed by a recommendation of the days of the week by Kgasa (1976) in the front matter of the Setswana dictionary. His list is a purist approach of avoiding borrowings from Afrikaans as he says:

> Malatsi a beke (tshipi) a ka bidiwa ka Setswana ka motlhofo go sena Sekgowa le fa e le Seburu (Kgasa 1976: front-matter).
>
> [Days of the week can be referred to easily without resorting to English or Afrikaans (*translation mine*)].

In the above quotation Kgasa is at pains in shrugging off borrowings but even the very Setswana sentence he uses to shun Afrikaans, has at least two borrowings from Afrikaans. These are *beke* 'week' and *Seburu* from 'Boer'.

Additionally, Kgasa rejected certain names of days of the week in standard Setswana such as *Matlhatso* which he considered to be religiously insulting to others. He objected that:

> Fa malatsi a bidiwa jaana ga gona nyenyafatso ya tumelo ya ba bangwe ka lefoko la Matlhatso jaaka go ntse gompieno (Kgasa, ibid)
>
> When the days of the week are referred to this way (in the way he suggested) there is no condescension of other people's faith with the term Matlhatso (Saturday) as it is today (*translation mine*).

Matlhatso is a noun derived from *tlhatswa* 'wash' and Kgasa may have perceived the name to be offensive to the Seventh Day Adventists (SDA) who consider Saturday as a day of rest and not for manual labour such as washing. Kgasa also objected to the use of the name *Mosupologo*:

> Lefoko la Mosupologo ga le utlwale ka gobo (sic, *go bo*) tota beke e a bo e sa robala mo e reng letsatsi le le salang Lantlha morago le bo le bidiwa Mosupologo jaaka ekete beke e a supologo (sic, *supologa*) (Kgasa, 1976: front matter).
>
> The word Mosupologo does not make sense because a week is not asleep, such that the day after Sunday should be called Mosupologo as if a week rises from dust (*translation mine*).

Kgasa understood that the noun Mosupologo is derived from the verb *supologa* 'rise from dust' and he found this inaccurate to refer to a day at the beginning of the week. But he was too late; the word had caught on and his recommendation never gained currency. His suggestion only jumbles the names of days of the week resulting with Monday called Tuesday (see Table 16). This failed attempt by Kgasa approximates Churchward's (1959) inventions of loan words in his dictionary (see Lichtenberk, 2003: 394).

What is surprising concerning Kgasa's recommendations is that Setswana authors before him did not share his views. For instance, Sandilands (1953: 153) days of the week are dissimilar to Kgasa's recommendations:

Table 17: Sandiland's rendering of days of the week

Setswana	English
Lamoréna[13], Tshipi	Sunday
Mantaga, Mosupologò	Monday
Lwabobedi	Tuesday
Lwaboraro	Wednesday
Lwabonè	Thursday
Lwabotlhano	Friday
Matlhatsò, Maapèò[14], Satertaga	Saturday

Although some of the terms used by Sandilands have since gone out of usage, his rendering of days of the week is closer to the way Setswana is currently spoken compared to Kgasa's recommendations.

But of immediate relevance to this section also is what we list as Spoken/Common names of the week. The list includes borrowings *Mantaga/Mmantaga, Sateretaga,* and *Sontaga* from Afrikaans *Maandag, Saterdag* and *Sondag* respectively. Contrary to Lichtenberk's recommendations, excluding these borrowings from a Setswana dictionary would make it highly deficient since they are common in spoken language and increasingly used in the media, parliament and other domains of Setswana language use as illustrated in the concordance lines below.

Figure 3: Mantaga concordance lines

```
1      e jaaka ekete ke tsatsi la Sontaga. Mantaga mongwe le mongwe thupa e n
2      eng thata ka metlae, e leng Luzboy, Mantaga le Laboraro mongwe le mong
3      o ga a site go sita loso. E ne e le Mantaga thapama fa Motsei a tswale
4      wa sebining ya ga Motsei. E ne e le Mantaga mme nako e ka nna ya bosup
5      ka pampiri (di-mask). 45 Lenaneo la Mantaga - Std 4 Bana ba ithuta ka
6      e dilo tsa gago tsa go ya tirong ka Mantaga. , : Mosadi o o jaaka wena
7       se tima. Fa rraagwe a ya tirong ka Mantaga, a gakgamatswa ke fa sejan
8      olo ya gore o tl ya kwa teropong ka Mantaga a ye go reka dipampiri go
9      eleng ba ne ba tla boela Tembisa ka Mantaga thapama. Bana ba ga Daphne
10     ile phitlhong pele ga e sutisiwa ka Mantaga, Mogokgo wa sekolo se sego
11     sigo ka Satertaga le ka Sontaga. Ka Mantaga o ne a tshwarwa ke dipapal
12     go go itsise gore o tla simolola ka Mantaga. : Ke tla kgona go ya tiro
13     RONE: Erile Palamente e simolola ka Mantaga, T ona ya T emo-thuo, Dani
14     ng lengwe le lengwe, go simolola ka Mantaga go ya kwa go Labotlhano. B
15     teretaga le erne jaana: simolala ka Mantaga — Sateretaga 6 a.m. ke nak
16     e 4 se ka a itse go tla sekolong ka Mantaga. Re a bona jaaka mosetsana
17     ka rakana le Mosela kwa sekoleng ka Mantaga. Re ne re na le boikutlo b
18     senwa. Go tloga fa re ya gae mme ka Mantaga o tla mpolelela maina a di
```

[13] The use of this word to refer to Sunday has almost disappeared from Setswana use and may only be found amongst very few old speakers of Setswana, in very rare occasions.

[14] The use of this word to refer to Saturday is no longer in current Setswana usage.

```
19    ne e le mafelo a beke a maleele, ka Mantaga e ne e le letsatsi la boik
20    a go tlhatswa dikhai tsa Makgowa ka Mantaga le ka Labobedi mme a be a
```

Figure 4: *Sontaga* concordance lines

```
1     ne yo o neng a tshaba go lema yole. Sontaga mongwe le mongwe bana ba d
2     botsa Mmadisenke mo tshokologong ya Sontaga ba robile sogo tno phaposi
3     esele! Ga ke tshoswe ke modumedi wa Sontaga fela. Mo bekeng re a tshwa
4     " Ga bua Kepaletswe monyebo e le wa Sontaga, "Dumela Kepaletswe. Ke en
5     tshameko ya bosheng, bogolo jang wa Sontaga mme re ka nametsega re le
6      tshipi. Ke t/aa go bona ka Tshipi (Sontaga). Ba tlaa goroga tshipi(be
7      sa ga Mosela mo nokeng e Tshetlha. Sontaga e e latelang go ne ga bewa
8     na f~a a laela. "Ke tla go tshakela Sontaga se se tlang ..." A bua a
9     se golo mo re ja dilo,le dipina tsa sontaga , gape ke batla go bona ba
10     sonola sonolega sonolegile sonotse Sontaga sonya sopaladitse sopalala
11    mo sonobolomo sonobolomo sonobolomo Sontaga Sontaga sontile soutile so
12     e telele. Kag~so o ile sekolong sa Sontaga le mme. Mme o farile Kagis
13    a Sontaga dipina. Bana ba sekolo sa Sontaga ba ntse mo ditilong. Ba op
14    tseboa 8.30 p jn. nako ya go robala Sontaga 7.00 a.m. nakoya go tsoga
15    majana a a lesome le bosupa. E rile Sontaga kefa lonyalo Iwa rona lo b
16     ffe yo Bham, o bula Sateretaga, le Sontaga tota o kgona go thusa bath
17    mo mafelong a beke, ka Matlhatso le Sontaga, fa a sa ya go bogela mots
18    eme leganbng. Go ne go le tsatsi la Sontaga, Ka nako tsa lesorne mo tl
19    lhela ka Sateretaga. Ka letsatsi la Sontaga ba lelwapa ba ne ba ya
20    a tlhomamiso ba tla tlhomamisiwa ka Sontaga. 6. Lokololа polelonolo e
```

Such names of the week could be marked in a dictionary as *common in spoken language*, or as *colloquial*. But it would be unsatisfactory not to list them in a Setswana dictionary just because a small set (of Afrikaans names of the week) from which they are derived, is not borrowed into the Setswana language in its entirety. Frequency here should be considered paramount.

The Setswana dictionaries have treated the different three borrowing in different ways. Brown (1925) does not enter *Mantaga*, *Sateretaga* and *Sontaga*. Kgasa (1976) enters *Mantaga* and not *Sateretaga* and *Sontaga*. Snyman et al (1990) include *Sontaga* and *Mmantaga* in the dictionary but leave out *Sateretaga*. Matumo (1993) does not enter *Mantaga*, *Sontaga* and *Sateretaga*. Kgasa and Tsonope (1998) enter *Sontaga* and not *Mantaga* and *Sateretaga*.

Word frequency lists are helpful in decisions of what to enter in a dictionary. Listing frequent borrowings such as *Sontaga*, *Mondaga* and *Sateretaga* and marking them as either colloquial, belonging to spoken language or as foreignisms would be a preferred approach.

Obviously the kind of dictionary being built would influence such decisions; whether it is monolingual or bilingual, intended for learners or for general use, or whether it is a dictionary of slang or not, primarily for encoding or decoding (e.g. academic use, which is a different case) and the number of pages a lexicographer has to work with.

Additionally, cases where certain terms, though known in the native language are rarely used in speech, but are replaced by borrowings and code-switchings, cannot be ignored (cf. Otlogetswe, 2006). This is particularly true for numerals where one finds sentences like, *O rekisitse dinamune di le* ten. "He sold ten oranges". *Mmiting o ka* ten *kamoso.* "The meeting is at ten tomorrow". In these examples, the speaker has chosen the English word *ten*, instead of the Setswana term *lesome/some*. The transcription of the term *ten* as either *ten* or *thênê*, as in the above examples, is based on the theoretical question of whether such a term has gained currency as an instance of borrowing or of code-switching. Are lexicographers to assume that such language usages do not exist in the language and that they do not have any relevance to dictionary compilation? Any answer to this question would lead to disagreements between lexicographers.

A similar pattern may be observed in days of the week with *Sateretaga* (Saturday)*, Sontaga* (Sunday)*, Mantaga* (Monday)*,* and *wikente* (weekend) being more colloquial and common in spoken language than in the written form while *Matlhatso* (Saturday)*, Tshipi* (Sunday)*, Mosupologo* (Monday) and *mafelo-a-beke* (weekend)*,* are common in written text, formal address and amongst the elderly. The stylistic information is significant particularly in dictionaries that attempt to achieve a fuller understanding of a word's meaning and usage. When both formal and informal terms are included in a dictionary, they may provide valuable stylistic information and may also be significant to future research as to when a word entered the language or when it changed its meanings.

This hopefully shows the importance of including greater occurrences of spoken text in a corpus since spoken language is used more in human communication and possesses unique characteristics not common in written language.

Next, the design of the two English corpora is considered.

4.5 Brown Corpus and BNC review

In Chapter 5 we discuss the Setswana corpus design and compilation. Before that we review two corpora which have been influential in English corpora analysis: The Brown Corpus and the BNC.

4.5.1 The Brown Corpus

Corpus linguists usually make reference to the *Brown University Standard Corpus of Present-Day American English*, commonly known as the Brown Corpus, (Francis and Kucera, 1964) as having pioneered research in corpus computational linguistics. The Brown Corpus was "significant not only because it was compiled for linguistic research, but also because it was compiled in the face of massive indifference if not outright hostility from those who espoused conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky" (Kennedy, 1998: 23).

The Brown Corpus was compiled by Nelson Francis and Henry Kucera in 1961. The corpus has over a million tokens of written text published in the USA in 1961. The Brown Corpus comprises 500 samples of about 2,000 tokens of continuous written English which approximate 1,014,300 tokens. Table 18 gives the text categories of the Brown Corpus and the proportions of different portions of the corpus.

Table 18: Structure of the Brown Corpus

Text type	**Proportion**
i. Informative Prose	**75%**
A. Press: Reportage e.g. Political, Sports, etc	8.8%
B. Press: Editorial e.g. personal, letters to ed., etc.	5.4%
C. Press: reviews e.g. books, music etc.	3.4%
D. Religion e.g. tracts, books, etc.	3.4%
E. Skills & Hobbies e.g. periodicals, books, etc	7.2%
F. Popular lore	9.6%

e.g. books, periodicals, etc	
G. Belles letters, biography, memoirs etc	15%
H. Miscellaneous e.g. government documents, industry reports, college catalogue, etc.	6%
I. Learned e.g. medicine, mathematics, law, etc.	16%
ii. Imaginative Prose	**25%**
J. General Fiction Novels and short stories	5.8%
K. Mystery and Detective Fiction Novels and short stories	4.8%
L. Science Fiction Novels and short stories	1.2%
M. Adventure and Western Fiction Novels and short stories	5.8%
N. Romance and Love Story Novels and short stories	5.8%
O. Humour Novels and essays, etc	1.6%

According to Kucera and Francis (1967: xvii) the samples were selected by "a method that makes it reasonably representative of current American English".

Ide and Macleod (2001: 274) argue that while the Brown Corpus has been extensively used for natural language processing work, its million words are not sufficient for today's large scale applications. For example, for tasks such as word sense disambiguation, many word senses are not represented, or they are represented so sparsely that meaningful statistics cannot be compiled. Similarly, many syntactic structures occur too infrequently to be significant. The Brown Corpus is also far too small to be used for computing the bigram and trigram probabilities that are necessary for training language models used in a variety of applications such as speech recognition. Fillmore et al. (1998: 966) have also found the Brown corpus to be "too small to provide adequately large samples for the purposes of lexicon construction."

Furthermore, the Brown Corpus, while balanced for different written genres, contains no spoken English data. Ide and Macleod (2001) lament the fact that while the 100 million words of the BNC provide a large-scale resource and include spoken language data; it is not representative of American English. As a result, there is no adequate large corpus of American English available to North American researchers for use in natural language and speech recognition work. Ide and Macleod (2001), because of this lack have argued that there is a need for a corpus of American English that is similar to the

British National Corpus. The project to compile the American National Corpus comparable to the BNC is detailed in Ide et al. (2002). They have shown that there are significant lexical and syntactic differences between British and American English. They point to the well-known variations such as: "at the weekend" (Br.) vs. "on the weekend" (U.S.), "fight (or protest) against <something>" (Br.) vs. "fight (or protest) <something>" (U.S.), "in hospital" (Br.) vs. "in the hospital (U.S.), "Smith, aged 36,…" (Br.) vs. "Smith, age 36…" (U.S.), "Monday to Wednesday inclusive" (Br.) vs. "Monday through Wednesday" (U.S.), "one hundred and one" (Br.) vs. "one hundred one" (U.S.), etc. Also, in British English, collective nouns like committee", "party", and "police" have either singular or plural agreement of verb, pronouns, and possessives, which is not true of American English.

Rayson and Garside report that the Brown corpus has been used in one of the largest comparative studies of the one million words of the American English (the Brown corpus) with one million words of British English (LOB corpus) by Hofland and Johansson. (1982). They also report on Yule's (1944) coefficient measurement which showed the relative frequency in the two corpora. Kilgarriff (1997a) used the Brown corpus to measure corpus homogeneity. The Brown corpus has also been studied for the abstraction of collocations. It has been found that the Brown Corpus has only two instances of "cups of coffee", five of "for good" and seven of "as always" (Kjellmer, 1994a).

The Brown corpus has therefore been a useful resource for linguistic research. However as has been seen, it was just too small for studies which needed large corpora. One corpus which was compiled to respond to this need is the British National Corpus.

4.5.2 The BNC review

The BNC is a 100 million-word corpus of written and spoken language from a variety of sources, designed to represent a wide cross-spectrum of current British English. The corpus "contains just over 4,000 texts" (Aston, 2001: 73). It was compiled by by a consortium of dictionary publishers and academic researchers between 1990 and 1994. These included the Oxford University Press, Longman Group Ltd, Chambers Harrap, Unit of Computer research on the English Language (Lancaster University), Oxford University Computing Services, and the British Library Research and Development Department. Ninety percent of the BNC are written texts while 10% of the BNC is transcribed spoken text.

The BNC compilation was funded over three years with a budget of over GBP 1.5 million. The project was funded by the commercial partners, the Science and Engineering Council (now EPSRC) and the DTI under the Joint Framework for Information Technology (JFIT) programme. Additional support was provided by the British Library and the British Academy (see the BNC website: http://www.natcorp.ox.ac.uk/).

4.5.2.1 The BNC design criteria

Since the BNC was compiled so that generalizations could be made on the British English it was crucial that varieties that existed in the British English be represented in the corpus. The BNC was therefore built by sampling materials from across the language with respect to explicit design criteria rather than basing the collection of texts on their availability. Burnard notes that,

> The objective was to define a stratified sample according to stated criteria, so that while no-one could reasonably claim that the corpus was statisticxally representative of the whole language in terms either of production or reception, at least the corpus would represent the degree of variability known to exist along certain specific dimensions, such as mode of production (speech or writing); medium (book, newspaper, etc.); domain (imaginative, scientific, leisure, etc.); social context (formal, informal, business, etc) and so on (Burnard, 2002: 21).

The BNC design criteria specify a range of text characteristics and proportions for the material to be collected (see Atkins, 1992). Below we briefly look at both the written and spoken language design criteria of the BNC.

4.5.2.2 The BNC written component

Ninety percent (89,740,544 words) of the BNC is written texts that were classified into two principal parallel categorisations of:

a. *domain* (i.e., subject matter, divided into nine classes, viz., imaginative; arts; belief and thought; commerce; leisure; natural science; applied science; social science; world affairs: from 146 to 527 texts in each), and

b. *medium* (five classes, viz., book; periodical; miscellaneous published; published; to-be-spoken: from 35 to 1,414 texts in each). All the texts were selected on the basis of a publication period, marked as *time* in the corpus (Aston, 2001: 73).

The written part includes extracts from regional and national newspapers, specialist periodicals and journals for different ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.

The criterion of *domain* refers to the content-type of the text; *time* refers to the period of text production, while *medium* refers to the type of text publication, as in newspaper or book. Table 19 summarises the contents of the three criteria (see Aston and Burnard, 1998: 28-33).

Table 19: The BNC written components

Domain	%
Imaginative	21.91
Arts	8.08
Belief and thought	3.40
Commerce and finance	7.93
Leisure	11.13
Natural and pure science	4.18
Applied Science	8.21
Social Science	14.80
World Affairs	18.39
Unclassified	1.93
Time	%
1960-1974	2.26
1975-1993	89.23
Unclassified	8.49
Medium	%
Book	58.58
Periodical	31.08
Misc. published	4.38
Misc. unpublished	4.00
To-be-spoken	1.52
Unclassified	0.40

4.5.2.3 The BNC spoken component

The design of the spoken component of the BNC adopted a two-part approach: demographic and context-governed. The demographic approach employed demographic parameters to sample everyday speech of the British English speakers in the United Kingdom. The context-governed approach attempted to cover the full range of linguistic variation found in spoken language using a typology based on four contextual categories: educational (lectures, news broadcasts etc), business (sales demonstrations, union meetings etc), public/institutional (sermons, political speeches etc) and leisure (sports commentaries, radio phone-ins etc) (Crowdy, 1994). The demographic component, on the other hand, comprises recordings of 124 volunteers from four different social classes, male and female, different age groups and various geographical regions.

The spoken component constitutes 10% (10,365,464 words) of the BNC. For the spoken component, a first distinction was between "demographic" (conversations: 153 texts) versus "context-governed" (speech recorded in particular types of setting: 757 texts), and the "context-governed" component was further divided according to the nature of the setting (educational/informative; business; public/institutional; leisure: from 131 to 262 texts in each), paralleled by a monologue/dialogue distinction (40%/60%) (Aston, 2001: 73). Table 20 summarises the divisions in the corpus. It covers both the demographic and context-governed components and the context-governed component structure.

Table 20: The BNC spoken components

Context-governed	**%**
Leisure	23.71
Institutional	21.86
Business	21.47
Educational and Informative	20.56
Unclassified	12.38
Region	**%**
South	45.61
North	25.43
Midlands	23.33
Unclassified	05.61
Interaction type	**%**
Dialogue	74.87

Monologue	18.64
Unclassified	06.48

The value of compiling such a stratified corpus was to try and capture the varieties of modern British English from the 60s until the early 90s. It was designed to characterise contemporary British English “in its various social and generic uses” (Aston and Burnard, 1998: 28). Such linguistic variability was crucial for the corpus so that authoritative generalisations about the language could be made confidently. This need for compiling representative corpora from which generalisations could be made and on which hypothesis could be tested is expressed by Renouf thus:

> When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested. The first step towards achieving this is to define a whole of which the corpus is to be sampled (Renouf, 1987: 2).

The BNC has been useful for a wide variety of language research purposes including dictionary compilation of the *Longman Dictionary of Contemporary English* (3rd edition) (Summers, 1995), *Oxford Advanced Learner’s Dictionary of Current English* (Hornby, 1996), *Longman Essential Activator* (1997) and *The New Oxford Dictionary of English* (Pearsall, 1998). The BNC “was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography” (Kilgarriff, 2001: 342).

Leech et al. (1997) explored the social differentiation in the use of English vocabulary in the BNC while Čermák and Kren (2005) compare its composition with that of Czech National Corpus. Rayson et al., (1997) undertake selective quantitative analyses of the demographically-sampled spoken English component of the British National Corpus. They compared the vocabulary of speakers according to gender, age and social group. The BNC has also inspired the compilation of other corpora such as the American National Corpus (Fillmore et at., 1998), the Russian Reference Corpus (Sharoff, 2004) and the Czech National Corpus (Čermák, 1997).

4.6 The exploration of both corpora

After a corpus has been compiled, "lexicographers need the skills and/or the software to navigate through sometimes huge numbers of corpus instances" (Kilgarriff, 2000: 109). However it has been found that there is a lack of tools for corpus-based lexicography, especially in relating corpus observations to dictionary entries (Heid, 1994; Simons, 1998). Confronted with huge amounts of data, researchers need statistical and computational methods to query it in meaningful ways. Such mastery has been demonstrated by Francis and Kučera (1982) in analysing the 1 million Brown Corpus of American English. They calculated the frequency lists of different word forms and the coefficient of their usage. A similar 1 million word-corpus was built at the University of Lancaster called The Lancaster-Oslo/Bergen Corpus (or the LOB corpus). It had a similar structure to the Brown Corpus but comprised British English (Johansson and Hofland, 1989). Johansson and Hofland did a study of the word frequencies on this corpus to determine the most frequent words. Frequency of usage is crucial to lemmatisation since it guides the lexicographer in determining a headword list. Research on the BNC (Leech et al., 2001) has been attempted involving sophisticated statistics to rank frequency lists of grammatical word classes of the whole corpus, spoken versus written text, and determining distinctiveness of the grammatical word classes of spoken versus written text. Rayson et al. (2002) have analysed the relationship between part of speech frequencies and text typology in the BNC. Levin et al. have used the BNC extensively to demonstrate the role corpus data has in lexical research and the development of a theory that explains and predicts word behaviour. Their research explored the verbs of sound. Other researchers have attempted to assess methodologies of determining which words are particularly characteristic of a text. Kilgarriff (1996) used the BNC to compare the chi-square test, Mann-Whitney ranks test, the t-test, Mutual Information statistic (Church and Hanks, 1989), log-likelihood (Dunning, 1993), poisson mixtures, adjusted frequencies, content analysis (Wilson and Rayson, 1993) and Biber's (1988, 1995) Multi-dimentional analysis in determining which statistical approaches are best suited to identifying words that are characteristic of a text. In the development of this thesis we will explore different statistical approaches to measure similarities and differences in corpus components.

These statistical and computational advancements of querying a corpus are characteristic of developments in research in the English language. Such studies have not been attempted in Setswana.

4.7 Conclusion

In this chapter an attempt has been made to show that, while corpus research stands as one of the most useful approaches to language research, particularly lexicography, in that it can speedily offer information for addressing language related issues and problems, a critical look at the process of corpus construction would help us determine if generalisations drawn from its results should be trusted as true reflection of language use. While corpus linguists are fairly in agreement about the inclusion of language varieties in a corpus, there is still a lack of clarity concerning whether a language population can be known and sampled in all of its varieties. In sampling such varieties, it is not clear how much of each variety is to be sampled. However this has not restricted lexical research to argue that "Corpora like the BNC are designed to provide sample data from which to infer generalisations about the language as a whole, or about particular broad categories of texts…" (Aston, 2001: 75). There are still differences on what it means for a corpus to be balanced and representative of a language from which it was abstracted.

The lack of spoken language and language varieties in many corpora stands as their greatest limitations. This is because the recording and transcription of spoken language is expensive and time-consuming. Communities such as the ones found in many African states face unique challenges to corpus compilation in that their languages are not used in various domains such as: academic writing, media, government and official communication, making text in these domains almost impossible to find. Since automatic transcription is as yet an unsolved problem, it means that attempts of building large corpora of spoken language may remain impossible for some time. The kind of corpus that compilers end up with is therefore the one characterised by Kilgarriff as

> …a corpus which will never be beyond challenge at a theoretical level, but which does nevertheless allow us to address with a degree of objectivity some central questions about the language, where before we could only speculate Kilgarriff (1997: 137).

We have also looked at two corpora, the Brown and the British National Corpus; the former with only a million words, and the later with 100 million words. The two corpora were built about 30 years apart; the Brown Corpus in the 60s and BNC in the 90s. We have inspected their internal structure and revealed that both corpora include samples from different domains to attempt a balanced representativeness of language as used. Both corpora were revolutionary for their times. The Brown Corpus was compiled at the time when hostility was high against impericism, while the BNC is unique for its size and variability. It is through building and querying balanced corpora (Kennedy, 1998; Ooi, 1998: 29) such as the two corpora through advanced statistical and computational approaches that a detailed analysis of a language could be achieved.