Chapter 8

Conclusion and future work

This research set out to try and determine how a Setswana corpus could be compiled and structured as a balanced and representative entity through both quantitative and qualitative means in order for it to be "better suited" for lexicography. The aim was to determine whether a corpus compiled with texts from various text types was better suited for lexicography or whether similar results could be attained through corpora compiled with texts from few or a single text type.

To test the aims, a Setswana corpus of over 13 million tokens was compiled. The compilation of this corpus has been discussed in Chapter 5.

Chapter 1 positioned this thesis squarely within the scope of corpus design with specific application to the lexicography of the Setswana language. The chapter outlined the goals and aims of the study, and the methodologies employed. The chapter concludes by taking a panoramic overview of the whole study with an exposition of chapters.

Chapter 2 explored the Setswana language situation particularly in Botswana. Multilingualism in Botswana was discussed and it was shown that Setswana is spoken by the majority of the Botswana population (about 80%). Despite the population's multicultural composition, it has been established that only two languages, Setswana and English, occupy a dominant position in the educational system (Mooko, 2004: 181/2). It was also established that English remains the official language and a language of considerable prestige, while Setswana is the national language and the country's lingua franca. Other Botswana languages apart from Setswana and English have no official status in Botswana (Molosiwa, 2004: 6) and remain excluded from functioning as mediums of instruction and use in the media (both broadcast and print, save for Ikalanga which is used minimally in *Mmegi* Newspaper insert, *Naledi*), parliament, or in any public domain to communicate government policy. It has been shown that while minority languages are in general marginalised from any official

function, in regions where minority languages are the regionally dominant languages, the minority language is usually used in official roles, like communicating with the chief or nurse (Hasselbring et. al. 2001: 32-33). The appraisal of the language situation in Botswana mapped out the complex language use and varieties of Setswana.

We also traced the history of Setswana lexicography and language research dating to the early missionary period and situated them within missionary literacy programs amongst the Batswana. Developments in corpus and computational models have been reviewed and it has been shown how they have affected dictionary compilation in Setswana. We have illustrated how the Setswana language could benefit from developments in corpora, corpus querying software (CQS) to produce frequency lists, concordances, and keyword analysis. By outlining Botswana's sociolinguistic situation, Chapter 2 established a foundation for Chapter 5, which outlines the design and compilation of the Setswana corpus.

Chapter 3 explored the theoretical issues related to corpus design. It surveyed various definitions of "corpus". The following findings were established in the definitions discussed:

- Corpora are usually "sufficiently large" for the research they have been compiled for.
- Corpora are collections of running texts. They are not just lists of words but rather chunks of texts like chapters of books, entire books, or transcribed speech.
- Corpora are compiled for some linguistic research.
- Because of their massive size, corpora are stored in computers because of the computer's storage and processing power. Computers are "good at recall, people are good at precision; that is, computers are good at finding a large set of possibilities, people are good judges of which possibilities are appropriate" (Kilgarriff, 2003: 1). They can also be used interactively, allowing the human analyst to make difficult linguistic judgements.

It was also shown how the Web, with its billions of words, has revolutionised the compilation of corpora. The Web was seen as providing a cheap route to corpus compilation (De Schryver, 2002). The benefits of Web text were revisited and demonstrated in Chapter 5 which details the compilation of half a million tokens through the use of a Web crawler.

Keyness and frequency profiling were introduced in Chapter 3 and later used in Chapter 6 and 7. Baroni (2006: 1) has observed that "The frequency of words and other linguistic units play a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language."

In Chapter 4 we discussed issues which arise in corpus design as they relate to lexicography. Corpus design is at the heart of this study since corpus design and compilation determine the quality of what could be extracted from a corpus. Balance and representativeness have been discussed and found to still be areas of great contestation in corpus design and compilation. It has also been shown that what constitutes balanced and representative corpora still remains controversial. The sampling of genre quantities for a corpus is still largely unresolved. The general consensus in the literature is that a corpus must capture the language varieties of a population from which a sample is taken, which reflects how that particular language community uses language. Such a goal is important since many corpus linguists hope to generalise the results of corpora analysis to the general language community from which the samples have been abstracted.

Sinclair (2004) has argued that the complicating factor in compiling balanced and representative corpora is that language is an "unlimitable phenomena". A quest to quantify language usually results in general estimates. Although language is an unlimitable phenomenon, that has not obstructed corpus researchers from arguing for sampling different linguistic varieties for both quantitative and qualitative study. The challenge for corpus linguists and lexicographers is to identify the language varieties of the language under study and ensure that they are represented in a corpus.

It has also been demonstrated that a corpus can contain simple raw text or it can be enriched with linguistic information which will enhance information extraction. Tagged corpora, it was argued, are useful in the development of disambiguation rules and the facilitation of automatic and semi-automatic syntactic analysis in corpus linguistic. Tagged corpora have also been found to be highly useful in the generation of Word Sketches "…one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al., 2004: 105).

In Chapter 4 the importance of spoken language in corpus design and compilation is discussed. Sinclair (2004) shows 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." However in none of the large corpora like the LOB, Brown Corpus, BNC and the Bank of English does the percentage of spoken text exceed that of written text. This state of corpora has been criticised by Sinclair (2004) thus: "most general corpora of today are badly balanced because they do not have nearly enough spoken language in them."

In the debate on spoken language inclusion in a corpus, Biber (1994) has argued that to have greater percentages of spoken language in a corpus is not linguistically interesting since it leads to corpus homogeneity. He has argued that what corpus compilers should aim for should be stratified corpora that capture the linguistic variability of the language community and not proportionally-compiled corpora.

Chapter 4 also discussed the importance of spoken language by illustrating what would be lost if a corpus lacked transcribed speech. Such losses would include borrowings and code-switching which are linguistically interesting. Setswana data has revealed that spoken Setswana has high concentrations of borrowings from English and Afrikaans and instances of code-switching. Chapter 4 concluded by reviewing the BNC and the Brown corpus. Their internal structures were studied to reveal how they compare with the Setswana corpus compiled for this study. Both corpora include samples from different domains to attempt a representativeness of the English language as used. Their analysis provided a base for Chapter 5 where we outlined the design and compilation of the Setswana corpus.

Chapter 5 mapped out the compilation of the Setswana corpus which has been used for experiments in this study. It is about 13 million words and covers texts from different varieties of Setswana language use including, novels, plays, newspapers, grammar books, spoken language covering court transcripts, call-in programs, television debates, funeral services, classroom interaction and sermons. The corpus, like many corpora, has large sections of written language and smaller sections of transcribed spoken language. Ninety four percent of the corpus is the written component while the spoken component occupies 6%. Its design and compilation was influenced largely by the structure of the BNC. The corpus is significant for the experiments which have been conducted in this thesis as Dash and Chaudhuri (2000: 188) have observed that "the potentiality of a well-designed corpus is immense as it provides an empirical basis for language description"

Keywords for Science and Technology, Politics, Poetry, Plays, Grammar, Arts and

Culture, Religious and Hansard text and interviews text from spoken language were calculated in Chapter 6. The statistical analyses were conducted through the use of WordSmith Tools. The Log Likelihood (G2) test was used to calculate keyness since it is considered better than the chi-square test of significance particularly when contrasting long texts or where one may have to deal with low counts of less than 5. We were here following Kilgarriff (2001: 105) who argued that "G2 is a mathematically well-grounded and accurate measure of surprisingness," and that "it corresponds reasonably well to human judgments of distinctiveness." The results of the experiment were presented as the top 100 keywords from each text type. Through keyword analysis it was found that different text types generated different keywords that were particular to them. Thus terms key to the following texts were abstracted: Religious, Science and technology, Politics, Poetry, Plays, Grammar, Arts and culture, Chat-site, News, Sport, Call-in, Face to face dialogue, Education, Hansard, and Interview and Open-radio programming.

The results of the experiments revealed that different text types contribute different keywords that are unique to them. The finding is significant to corpus design for it lends support to the inclusion of texts from a variety of text types in a corpus. Since text types are characterised by unique words which are key to them, a corpus comprising texts from the varieties of a language will be richer in its representation of a language.

Chapter 7 measured how for each text type the number of types grew with every additional 10,000 tokens. Our aim with this experiment was to investigate whether different text types' vocabularies vary at comparable token points. It was found out that taking measurements at 10,000 token-chunk intervals is sensitive to the order in which texts (i.e. 10,000-word corpus chunks) are placed or ordered. The ordering of the 10,000 token-chunks raised unique challenges in that every experiment repetition with a random order of the 10,000 token chunks gave different results dependent on which 10,000 token-chunk was analysed first.

The 10,000 token-chunks were randomised for every measurement taken and the experiment iterated five times to resolve the bias of sequence. A mean was computed so that comparisons could be made between text types using an average that summarised the results.

For additional experiments the written part of the Setswana corpus was divided into 10 text types of Poetry, Grammar, Chat-site, Plays, Prose, Science, Politics, Business,

Religious, and Newspaper texts. The spoken subcorpus was segmented into two parts: Hansard and Call-in (comprising interview, call-in text and open-radio programming treated as a single unit). Additionally, the 12 text types of spoken and written text were divided into three random groups of A, B and C with each group having four text types. These were labelled using the initial three letters of each text type found in each group, thus: POEGRACHAPLA (Poetry, Grammar, Chat-site and Plays), PRONEWHANCAL (Prose, Newspaper Hansard and Call-in) and SCIPOLBUSREL (Science, Politics, Business and Religious). The experiment compared subcorpora containing unrelated texts with equal-sized subcorpora containing text from a single genre. The TTR measure at comparable points for both texts was computed. The assumption was that combining text from a variety of sources (as one might do in corpus compilation) would give a higher TTR at comparable points compared to that of an equal-sized subcorpus with a single text type.

Randomly sampled fifty 10,000 token-chunks were taken from the 4 text types in each group. The final number of text types came to 15 including the three "constructed text types." Measuring word types of text types at 10,000 tokens intervals revealed that Poetry, PRONEWHANCAL and SCIPOLBUSREL had the largest number of word types in all text types measured. It was concluded that the high levels of types in poetry offer support to the view that poetic language is characterised by high lexical density. The high levels of types in PRONEWHANCAL and SCIPOLBUSREL led to the conclusion that corpora comprising a variety of text types have a higher number of types, than those compiled from a single text type.

The most frequent 100 words of various text types were compared with the top 100 words of the whole corpus. It was found that it was not enough to have a corpus which has a large variety of text types for one to generate large numbers of word types. Rather, it was crucial that the individual text types that comprise a corpus should individually have higher levels of word types themselves.

Simple consistency analysis (SCA) which calculates dispersion or word-spread in corpora was explored in frequency analysis. SCA demonstrated whether a high frequency word was high in the frequency list because it occurred in many of the text samples or whether it was because it was used frequently in a few texts. The SCA calculation computes words which recur consistently in texts and orders them on the basis of their spread. SCA was compared to raw frequency calculations of the most frequent words in the corpus. It was found that the compilation of headword lists would be enhanced by combining SCA calculations and raw frequency lists.

Raw frequencies were then used in the comparison of the most frequent 100 words of different text types so that our results could be comparable to those of other wordlists of other corpora such as the BNC. The most frequent 100 words of the spoken text and the most frequent 100 words of the written part of the corpus were compared with the most frequent 100 words of the entire corpus. It was found that 81 of the top 100 words of the written component of the corpus are found amongst the most frequent 100 words of the whole corpus. On the other hand, only 71 words of the top 100 words of the spoken component were found amongst the most frequent 100 words in the entire corpus. The written component of the corpus is larger and is much more diverse in the kind of texts it comprises while the spoken component is smaller and limited in its text diversity. This may explain the differences between the results of the spoken and written corpus component. Although 94% of the whole corpus is written material, still all of the top 100 in the whole corpus are not found amongst the most frequent 100 words of the written component. This fact suggests that even with its great diversity and size, written language alone is not adequate to make a representative corpus. There is a need for spoken material to be included in the corpus.

The Setswana experiment results of the written and spoken language were compared to those of experiments on the BNC. It was found that 71 of the Setswana spoken subcorpus' most frequent 100 words were amongst the most frequent 100 Setswana words. Eighty one of most frequent 100 written words were found amongst the most frequent 100 words of the complete Setswana corpus. The BNC, on the other hand had 97 of the top 100 words of the written component of the corpus in the most frequent 100 words of the whole corpus and 72 words of the top 100 words of the spoken component. The results of the two experiments were found to be similar and the Setswana corpus components comparable to those of the BNC since the BNC has 90% written material and 10% transcribed speech while the Setswana corpus has 94% written material and 6% transcribed speech.

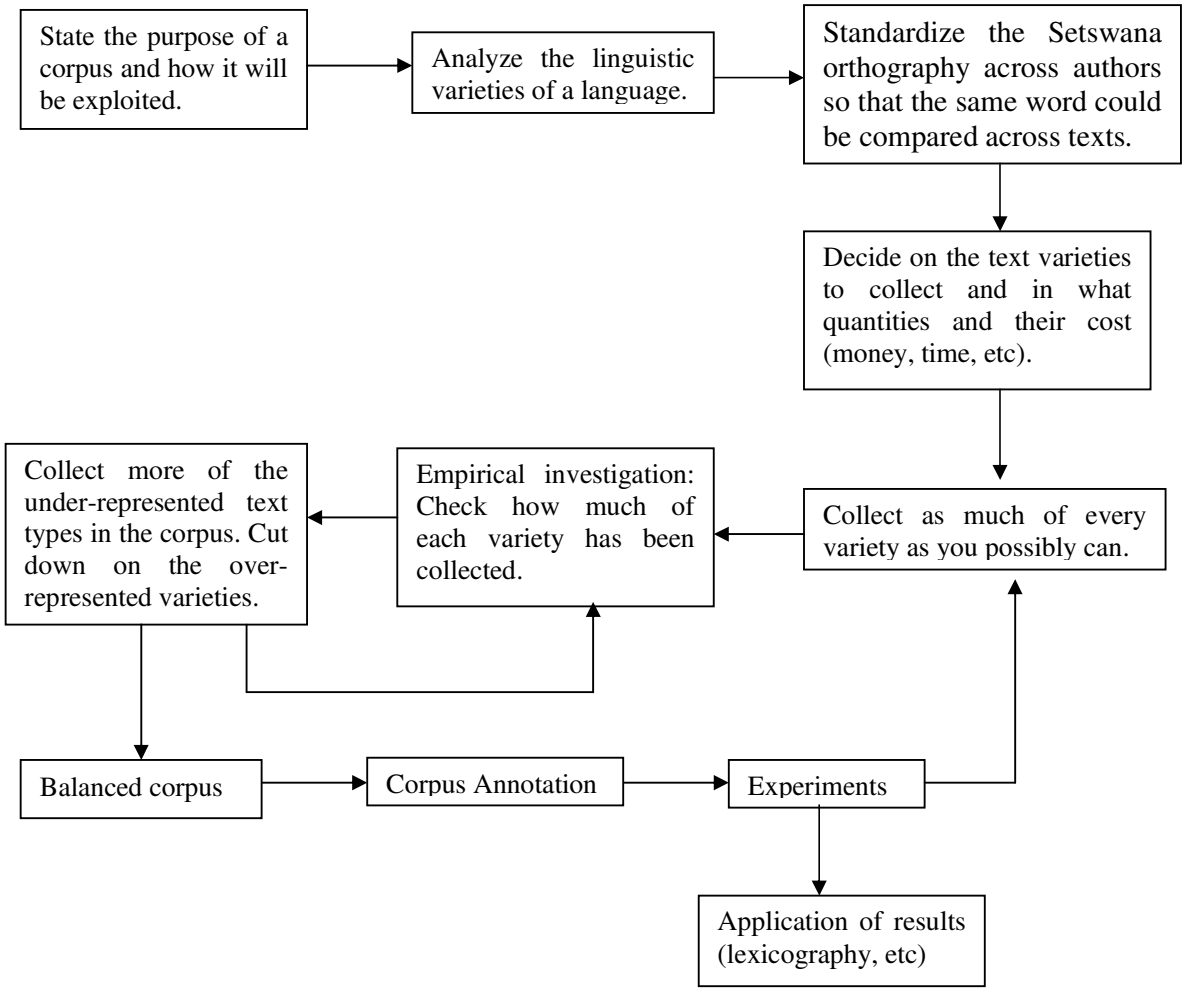
To further test whether text type diversity was crucial to the kind of words selected for inclusion in a dictionary, two 5,000-word list chunks were compared. The first chunk simulated an opportunistic corpus with its text type limitations. It was derived exclusively from prose text. Prose text was chosen since much of the readily available text in African languages is of a prose type. The majority of such text comprises novels. The most frequent 5,000 words from prose text were compared with 5,000 words from the following text types: Newspaper text, Religious, Chat-site, Hansard,

Poetry, Prose, Politics, Science, Call-in and Business. The comparison of the two 5,000-word lists was to determine which of the two lists covered a broad scope of linguistic varieties similar to those found in Setswana varieties. Christian terms, terms common in traditional Setswana beliefs, grammatical terminology, business terms and vulgarities were tested for in the two wordlists.

The results of all the experiments consistently showed that the 5,000 wordlist compiled from a diversity of text types performed better that the 5,000 wordlist from prose text. The 5,000 wordlist compiled from a diversity of text types was found to have a large number of Christian terms, words common in traditional Setswana beliefs, grammatical terms, business words and vulgarities. The results suggest that an opportunistic corpus is largely unreliable as a source of dictionary material. The simulated opportunistic corpus consistently lacked words which were in the simulated wide-coverage corpus. The results therefore give support to broad text type coverage in corpora compilation as a reliable source of broad lexical coverage in dictionary compilation.

The experiments conducted in both Chapter 6 and 7 provide an overwhelming support for a diversity of text types in corpus compilation. The implications for corpus design are that stages that can be followed for a model development of a Setswana corpus can be proposed. The same model may be extended to languages similar to Setswana. Below we develop a corpus compilation schema which has been influenced by Biber's schematic representation of the corpus construction (Biber 1994: 400) and by the experiences drawn from this study. The schematic representation illustrates how the purpose to which a corpus is created and an understanding of language varieties that exist in a language inform the corpus model that guides the construction of balanced corpora. The schema is rendered in Figure 18.

Figure 18: Schematic representation of a balanced corpus construction



The schema portrays corpus compilation as a continuous process of collection of text and attempting to balance the different text types against each other. This schema shows that as more text is collected more balancing should be attempted. The collection of more text and subsequent corpus balancing should however not curtail the use of the corpus since more text should ideally be continuously collected particularly for languages like Setswana where texts are generally rare.

8.1 Future research and applications

This study has attempted to contribute to the body of research in corpus design for the Setswana language and languages in a similar position to Setswana. More has to be investigated concerning Setswana corpus design and use on language similar to Setswana. It was not this thesis' aim to attempt to cover everything in corpus design,

such a goal is unattainable. It is hoped that this study will generate debate and research on design and corpus use. More work still has to go into the POS tagging of Setswana corpus to maximise its exploitation. The corpus exploitation discussed in Chapter 4, such as in the case of Word Sketches reported by Kilgarriff (2004) would benefit Setswana research if the corpus was tagged.

Although we have built a 13 million corpus with a variety of text types for this thesis, more text still needs to be collected particularly, spoken language with its diverse varieties. The recording of Setswana dialects for inclusion in the corpus may also shed light on the different terms and sentence structures particular to regional dialects.

It is hoped that this study opens new doors to increased corpora study of Setswana. Genre studies on the basis of corpora evidence have not been attempted in the Setswana language, largely because of the lack of Setswana corpora that comprise different genres. Keyword analysis of text types in Chapter 6 and other experiments in Chapter 7 that explored the text type differences have resulted in fruitful findings which lexicographers, sociolinguistics and others linguists could benefit from. It is also hoped that the methodologies employed and findings of this research will all prove fruitful to other language researchers.

Dash and Chaudhuri (2000: 180) observe that "[p]eople in every branch of information science now realize that a corpus, as a sample of living language, can open up new horizons of study and research." It is therefore hoped that the 13 million Setswana corpus compiled during this study will be a resource for corpora investigation of different aspects of Setswana research beyond this study such as morphology, syntax and further investigations of text type variability.