Text Variability Measures in Corpus Design for Setswana Lexicography

Text Variability Measures in Corpus Design for Setswana Lexicography

By

Thapelo J. Otlogetswe

CAMBRIDGE SCHOLARS PUBLISHING

Text Variability Measures in Corpus Design for Setswana Lexicography,
by Thapelo J. Otlogetswe

This book first published 2011

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-2637-5, ISBN (13): 978-1-4438-2637-2

TABLE OF CONTENTS

List of Tables viii

List of Figures xi

Preface xii

Acknowledgements xiii

Abbreviations xv

Chapter One 1
Introduction
- 1.1 Background to the study
- 1.2 Statement of the research problem
- 1.3 Clarifying terms: genre, text type and varieties
- 1.4 Methodology
- 1.5 Aims of the study
- 1.6 Research goals
- 1.7 Exposition of chapters

Chapter Two 11
The Setswana Language
- 2.1 The Botswana language situation
- 2.2 The Setswana language
- 2.3 Setswana dialects
- 2.3.1 The village, cattlepost, lands and city language
- 2.4 Domains of Setswana language use
- 2.4.1 Education
- 2.5 Text categories
- 2.6 Challenges of multilingualism and diglossia
- 2.7 The poverty of data
- 2.8 Setswana language research
- 2.9 Conclusion

Chapter Three 27
Corpus Lexicography
3.1 Introduction
3.2 What is a corpus?
3.3 Web as corpus
3.4 Frequency profiling: frequency and type/token
3.4.1 Frequency counts
3.4.2 Type/token and problems of wordhood
3.5 Relevance of corpora to lexicography
3.6 Some pre-electronic frequency studies
3.7 Electronic-corpora studies
3.8 Keyword analysis
3.9 Business keywords
3.10 Concordance
3.11 A review of existing methods of headword list identification
3.12 A historical perspective of headword lists
3.13 Non-corpus dependant methods of dictionary compilation
3.14 Semantic domains
3.15 Corpus lexicography and Setswana dictionaries
3.16 Conclusion

Chapter Four 70
Issues in Corpus Design for Lexicography
4.1 Introduction
4.2 Balance and representativeness
4.3 Corpus annotation
4.4 Sample size
4.5 Brown Corpus and BNC review
4.6 The exploration of both corpora
4.7 Conclusion

Chapter Five 119
The Setswana Corpus Compilation
5.1 Introduction
5.2 The design strategy
5.3 Overall corpus statistics
5.4 The Zipfian distribution
5.5 Corpus components
5.6 The compilation of corpus components
5.7 Conclusion

Chapter Six 145
Measuring Text Type Diversity
6.1 Introduction
6.2 Keyword analysis
6.3 Conclusion to keyword analysis

Chapter Seven 194
Type/Token Measures of Corpus Chunks
7.1 Type/token measures
7.2 Text divisions for experiments
7.2.1 Newspaper Components type/token
7.3 Conclusion of type-token measurements
7.4 A comparison of the top 100 tokens
7.5 A direct comparison of Setswana spoken and written corpus components
7.6 Conclusion

Chapter Eight 265
Conclusion and Future Work
8.1 Future research and applications

Bibliography 275

Appendix 1: Proposed Subentries of *pelo* Headword 293

Appendix 2: Participation Consent Form 295

Appendix 3: Conversation Log 297

Appendix 4: Headteacher's Letter 299

Appendix 5: Accompanying Details for Classroom Recordings 301

Appendix 6: Letter to publishers asking for text 303

Appendix 7: BNC Part-of-speech Codes 305

Index 311

LIST OF TABLES

Table 1: Botswana's linguistic and ethnic structure
Table 2: Number of speakers of Botswana languages
Table 3: The Setswana text types rendered in the BNC style
Table 4: Some of Henry Salt's Setswana terms
Table 5: Top 20 words in the Setswana corpus ranked by word spread
Table 6: Top 20 words in the Setswana corpus ranked by word spread
Table 7: Top 100 Mokgosi sport tokens with functional words
Table 8: Mokgosi sport list's top 100 tokens without functional words
Table 9: Mokgosi top 100 sports keywords
Table 10: Mokgosi business keywords
Table 11: Corpus derived possible subentries of *pelo* entry
Table 12: Corpus derived possible subentries of *mpa* entry
Table 13: Corpus derived possible subentries of *molomo* entry
Table 14: Corpus derived possible subentries of *lonao/dinao* entry
Table 15: Corpus derived possible subentries of matlho entry
Table 16: Setswana days of the week
Table 17: Sandiland's rendering of days of the week
Table 18: Structure of the Brown Corpus
Table 19: The BNC written components
Table 20: The BNC spoken components
Table 21: Overall corpus statistics
Table 22: Top 20 Setswana tokens
Table 23: Top 1000 token-ranges and percentages in the whole Setswana corpus
Table 24: Top 20 Setswana tokens
Table 25: The corpus written and spoken components
Table 26: Spoken components statistics
Table 27: Overall statistics of the written subcorpus
Table 28: STTR measures of the written subcorpus
Table 29: Newspaper component statistics
Table 30: Prose component statistics
Table 31: A contingency table
Table 32: Science and technology keywords
Table 33: Politics text keywords
Table 34: South African Setswana politics terms and Botswana Setswana politics terms
Table 35: Poetry text keywords
Table 36: Plays text keywords
Table 37: Plays text keywords with names treated as metatext
Table 38: Grammar texts keywords

Table 39: Arts & culture text keywords
Table 40: Chat-site text keywords
Table 41: News text keywords
Table 42: Religious text keywords
Table 43: Call-in text keywords
Table 44: Face to face dialogue keywords
Table 45: Educational spoken text keywords
Table 46: Hansard spoken text keywords
Table 47: Interviews spoken text keywords
Table 48: Open radio programming keywords
Table 49: Religious spoken text keywords
Table 50: Sport spoken text keywords
Table 51: Possible SPORT candidates
Table 52: Newspaper types at 10,000 word tokens intervals
Table 53: A table of means for Newspaper types
Table 54: Newspaper type scores with mean and standard deviation scores
Table 55: Newspaper type scores with mean, critical value, standard deviation and confidence interval scores
Table 56: Written subcorpus text types
Table 57: Three divisions of text types
Table 58: Fifteen major corpus text types
Table 59: Poetry, Grammar, Chat-site, Plays, POEGRACHAPLA text types
Table 60: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types
Table 61: Science, Politics, Business, Religious and SCIPOLBUSREL types
Table 62: Newspaper components types
Table 63: Top 100 most frequent tokens in the whole corpus
Table 64: Top 100 words: Simple Consistency Analysis results
Table 65: Poetry, Grammar, Chat-site, Plays and POEGRACHAPLA
Table 66: Science, Politics, Business, Religious and SCIPOLBUSREL
Table 67: Prose, Hansard, Call-in, Newspaper and PRONEWHANCAL
Table 68: Comparison of written and spoken components to the whole corpus
Table 69: The BNC top 100 words of the whole corpus
Table 70: The BNC top 100 words of the written corpus component
Table 71: The BNC top 100 words of the context-governed spoken corpus
Table 72: The BNC top 100 words of the demographic spoken corpus
Table 73: The BNC top 100 words of the spoken part of the whole corpus
Table 74: Comparison of the top 100 words of the BNC against the top 100 words of the written and spoken subcorpora
Table 75: Comparison of BNC and Setswana
Table 76: Outstandingly frequent spoken language
Table 77: Outstandingly infrequent spoken tokens
Table 78: Top 100 tokens of Prose and Combined list
Table 79: Christian terms
Table 80: TSB terms
Table 81: Christian terms and their ranks on the two lists
Table 82: TSB terms and their ranks on the two lists

Table 83: Grammar terms and their position on the two lists
Table 84: Business terms and their rank on the two lists
Table 85: Vulgarities and their position on the two lists

LIST OF FIGURES

Figure 1: Concordance results of the word *pelo*
Figure 2: Word sketch for pray (v)
Figure 3: Mantaga concordance lines
Figure 4: *Sontaga* concordance lines
Figure 5: A rapid frequency decline in the top 100 words
Figure 6: Spoken and written language corpus components pie chart
Figure 7: Setswana corpus text types
Figure 8: Spoken components statistics
Figure 9: Newspaper text division
Figure 10: Newspaper types at 10,000 word tokens intervals
Figure 11: Prose, Grammar Chat-site, Plays and POEGRACHAPLA types
Figure 12: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types
Figure 13: Science, Politics, Business, Religious and SCIPOLBUSREL types
Figure 14: Newspaper components types
Figure 15: Comparison of the three overall top text types
Figure 16: Comparison of the three overall lowest text types
Figure 17: 5,000 words from a variety of sources
Figêure 18: Schematic representation of a balanced corpus construction

PREFACE

This book is about the design of a Setswana corpus for lexicography. While various corpora have been compiled and a variety of corpora-based researches attempted in African languages, no effort has been made towards corpus design. Additionally, although extensive analysis of the Setswana language has been done by missionaries, grammarians and linguists since the 1800s, none of such research is in corpus design. Most research has been largely on the grammatical study of the language.

The recent corpora research in African languages in general has been on the use of corpora for the compilation of dictionaries and little of it is in corpus design. Pioneers of this kind of corpora research in African languages are Prinsloo and De Schryver (1999), De Schryver and Prisloo (2000 and 2001) and Gouws and Prisloo (2005).

Because of a lack of research in corpora design particularly in African languages, this book is an attempt at filling that gap, especially for Setswana. It is hoped that the finding of this study will inspire similar designs in other languages comparable to Setswana.

We explore corpus design by focusing on measuring a variety of text types for lexical richness at comparable token points.

The study explores the question of whether a corpus compiled for lexicography must comprise a variety of texts drawn from different text types or whether the quality of retrieved information for lexicographic purposes from a corpus comprising diverse text varieties could be equally extracted from a corpus with a single text type. This study therefore determines whether linguistic variability is crucial in corpus design for lexicography.

ACKNOWLEDGEMENTS

This book is based on PhD research started at the Information Technology Research Institute (ITRI), University of Brighton, Brighton (UK), under the supervision of Adam Kilgarriff and Roger Evans. After about two years of research at Brighton, the study was moved to and later completed at the University of Pretoria, Pretoria (South Africa) under the supervision of Daan Prinsloo and Adam Kilgarriff. I am therefore indebted to many individuals I met during the initial stages of the study and to the two institutions for the resources and expertise they availed to me. In particular I wish to express my heartfelt gratitude to my supervisors:

- **Dr. Adam Kilgarriff.** I first met Dr Kilgarriff's research as a postgraduate student at the University of Oxford. Since then I wanted to be his student. It was been a wonderful and enriching experience to study under his excellent supervision.

- **Dr. Roger Evans.** Before my transfer from the University of Brighton (UK) to the University of Pretoria (SA), Roger Evans was one of my supervisors. I am exceedingly grateful for his guidance in the earlier stages of this study.

- **Prof. Daan Prinsloo.** I am grateful to Prof Prinsloo's exceptional supervisory leadership and patience with editing my work and for offering me access to some Setswana texts.

I also wish to thank the following:

.... colleagues at ITRI, particularly Ying Ling. ITRI provided me with the finest research atmosphere for my study to flourish earlier in my PhD.

… Steve Crowdy (Longman dictionaries, UK) who shared the documentation of the BNC spoken corpus design and transcription with me.

… Mike Scott who availed many statistical papers and clarification of some of the programming behind Wordsmith Tools 4.

.... I am indebted also to my sponsor, The University of Botswana for funding the larger part of my PhD.

.... I am equally grateful to the committee Overseas Research Students Awards Scheme (ORS) administered by Universities UK committee for selecting me for one of their awards. The award paid for part of my tuition at the University of Brighton. The award was given on a competitive basis to international postgraduate research students of outstanding merit and research potential.

.... Thanks to the following for availing the Setswana text for inclusion in the corpus

- Prof. D.J. Prinsloo, University of Pretoria.
- Botswana Macmillan.
- Botswana Parliament.
- *Mokgosi* newspaper.
- Many secondary schools whose identity we are not disclosing in the interest of anonymity.
- Department of Information and Broadcasting.
- Different Botswana government departments.
- Prof. Kevin Patrick Scannell of the Department of Mathematics and Computer Science, Saint Louis University for helping with harvesting Setswana text on the Web.
- Dr Elma Thekiso (University of Botswana) who was kind enough to give us her court transcriptions.
- Different families and individuals who allowed us to tape their conversations.
- Thanks to Motlhaleemang Ntebela for giving us text from the *Mmegi* newspaper.

ABBREVIATIONS

BNC	British National Corpus
CDIF	Corpus Development Interchange Format
CI	Confidence Interval
CLAWS	Constituent Likelihood Automatic Word-tagging System
COBUILD	Collins Birmingham University International Language Database
CQS	corpus querying software
DDP	Dictionary Development Process
HLT	Human Language Technology
HTML	Hyper-text mark-up language
JFIT	Joint Framework for Information Technology
KWIC	Key Word in Context
LDOCE	Longman Dictionary of Contemporary English
LMS	London Missionary Society
LOB	Lancaster/Oslo-Bergen Corpus
MD	Multi-Dimensional
MWE	multi-word expression
NLP	Natural Language Processing
OED	Oxford English Dictionary
POS	Part of speech
RNPE	Revised National Policy on Education
RRC	Russian Reference Corpus
SCA	Simple Consistency Analysis
SDA	Seventh Day Adventist
SGML	Standard Generalized Mark-up Language
SIL	Summer Institute of Linguistics
STTR	Standardized type/token ratio
TEI	Text Encoding Initiative
TSB	Traditional Setswana Beliefs
TTR	Type/token ratio
WWW	World Wide Web
XML	Extensible Mark-up Language

CHAPTER ONE

INTRODUCTION

1.1 Background to the study

This book is about corpus linguistics, precisely corpus design for lexicography (the science and art of dictionary compilation) as it relates to the Setswana language. The field of corpus linguistics is broad, covering areas such as grammatical studies, language education sociolinguistics, phonetics, phonology, stylistic analysis, dialectology and others (Kennedy, 1998). Corpus linguistics, particularly its application to lexicography is in its infancy in many African languages, particularly so in the language which is the focus of this book: the Setswana language. The larger body of Setswana research and that of many African languages covers broad linguistic areas such as language attitudes and use (Savage, 1990; Mooko, 2002; Bagwasi, 2003), language ecology (Andersson and Janson, 1997), grammar (Cole, 1955), syntax (Demuth and Johnson, 1989) phonology and phonetics (Jones and Plaatjie, 1916/1928; Mathangwane, 2002; Chebanne, 2002), and language literacy (Molosiwa, 2004).

Almost all of the studies mentioned in the preceding paragraph do not use corpora. Those that use corpus data are in the minority and relate to the use of corpora for lexicography. Amongst these are Prinsloo and Gouws (1995) Gouws and Prinsloo (1997) Prinsloo and De Schryver (1999) and Prinsloo (2004). Furthermore most research in corpora for the African languages is aimed at the compilation of corpora for lexicographic use and not in corpus design. This study focuses on Setswana corpus design whose output can serve a lexicographic purpose. Its findings and methodologies it is hoped would inspire similar designs in other African languages.

In corpus research in general, the focus has been placed on what researchers can retrieve from corpora, amongst these being frequency information, lemma lists, example sentences in dictionaries and concordance lines (De Schryver, 2002: 275/6). While there is nothing defective with such studies, what is lacking in the literature is detailed and in depth

research on corpus design particularly for African languages. The gap is particularly worrying in that the quality of corpus output is dependant on corpus design.

Few corpus designs have been documented. Francis and Kucera (1982) document the meticulous nature of the Brown Corpus design, while Crowdy (1991, 1993 and 1994) discusses in detail the sophistication of the British National Corpus spoken component compilation and Burnard (1995) outlines the design of the entire British National Corpus. On the basis of what has gone into such corpora, researchers are able to determine how valuable corpus output of such corpora is. In our research we have not found any study in corpus design which outlines the design of any corpus in African languages. This book's objective, as will be outlined below, in part is to fill this gap.

1.2 Statement of the research problem

Corpora use is not common in many dictionary projects in Africa languages, Setswana included. The larger body of research in corpora is on corpus usage and rarely in corpus design. There is no research that focuses on the design of Setswana language corpora.

At a practical lexicographic level, the production of dictionaries in various African languages has been very low particularly when compared with dictionary compilation in English by publishing houses such as Oxford University Press, Longman, Webster, COBUILD (The Collins Birmingham University International Language Database) and Chambers. For instance since 1875 less than ten Setswana dictionaries have been compiled. Three of these are monolingual dictionaries (Kgasa, 1976; Kgasa and Tsonope, 1998 and Dent, 1992), one is trilingual (Snyman et al., 1990), and three are bilingual (Brown, 1925, Matumo, 1993 and Créissels and Chebanne, 2000). More dictionaries could have been compiled considering that Setswana has official status in South Africa and it is Botswana's national language (and not its official language as Onibere et al. (2001: 503) claim). None of the Setswana dictionaries mentioned above used corpora save for Kgasa and Tsonope (1998).

At a theoretical level, several corpus design issues are still to be explored. The question of how corpora should be compiled as resource bases for lexicography is still to be sufficiently researched. There is therefore a need to measure how best to design corpora whose output will closely reflect the character of the varieties of Setswana as they are used. At the centre of this book, therefore, is the question: what kind of corpus is "better suited" for Setswana lexicography? The question translates into the

following issues:

1. Which text types exist in the Setswana language? In which contexts is the language used? These questions are significant since what we wish to establish is the language text types that could be added to the compilation of a corpus. Beyond that, experimentally we want to calculate and measure which words are typical of a text type.

2. The lack of structured corpora on which experiments can be conducted remains a huge problem for many languages. In many cases of African languages there are no corpora, and in cases where they exist, they are usually purely *opportunistic*; a simple gathering of whatever text exists without an attempt of representing language variability in the structure of the corpus. The question that needs addressing is therefore, how best to compile a corpus or corpora for Setswana lexicography but also for other Human Language Technology (HLT) and Natural Language Processing (NLP) purposes which capture the linguistic variability of the language. Additionally, what types of language components should go into the corpus composition and in what quantities? Finally, how can we empirically account for what constitutes corpora for lexicography in Setswana?

1.3 Clarifying terms: genre, text type and varieties

Before we proceed further in this book, it is important that we briefly define the terms: text types, genre and varieties which are sometimes used differently in the literature. We discuss how various scholars use the terms and how the terms are used in this study. Genre has been defined thus:

> …texts that have a similar set of purposes, mode of transmission and discourse properties (Roberts, 1998: 79).

> …a category assigned on the basis of **external** criteria such as intended audience, purpose, and activity type, that is, it refers to conventional, culturally recognised groupings of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the **internal** (linguistic) criteria forming the basis of text type categories (Lee, 2001: 38; emphasis in the original).

> Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic; they are assigned on the basis of use rather than on the basis of form (Biber, 1988: 170)

Bussmann defines text types

> ….a term from **text linguistics** for different classes of **texts**. Within the framework of a hierarchical text typology, text types are usually the most strongly specified class of texts (e.g. recipes, sermons, interviews), characterised by different internal and external features (Bussmann, 1996: 481/2).

He also defined linguistic variety as,

> … a generic term for a particular coherent form of language in which specific extralinguistic criteria can be used to define it as a variety. For example, a geographically defined variety is known as a **dialect**, a variety with a social basis as a **sociolect**, a functional variety as a jargon or a **sublanguage**, a situative variety as a **register** (Bussmann, 1996: 512).

> One way of making a distinction between *genre* and *text type* is to say that the former is based on external, non-linguistic, "traditional" criteria while the latter is based on the internal, linguistic characteristics of texts themselves. A *genre*, in this view, is defined as a category assigned on the basis of external criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features, which are, instead, the internal (linguistic) criteria forming the basis of *text type* categories (Lee, 2001: 38).

Lee also argues that genre and register overlap:

> The two terms *genre* and *register* are the most confusing, and are often used interchangeably, mainly because they overlap to some degree. One difference between the two is that *genre* tends to be associated more with the organisation of culture and social purposes around language and is tied more closely to considerations of ideology and power, whereas *register* is associated with the organisation of situation or immediate context. Some of the most elaborated ideas about genre and *register* can be found within the tradition of systemic functional grammar ((Lee, 2001: 41/42).

Some linguists make distinctions between genres, domains and text types, as in Lee (2001). In this book such distinctions are not applied, instead we use genre, text types and varieties inter-changeably to refer to linguistic variability in general. Our position is similar to that of Aston (2001: 73) who uses the term “the term "text type" as a neutral one which does not imply any specific theoretical stance” but rather in general to refer to linguistic variability.

1.4 Methodology

There is a large body of lexical research which deals with comparing different language varieties to measure language variation or describe lexical qualities of a subcorpus (Biber, 1993; Kilgarriff, 1996, 1997a; Leech et al., 2001; Sharoff, 2006). Other comparisons and measurements have been done at the level of corpora (Kilgarriff and Salkie, 1996) where corpora have been compared for similarity and homogeneity through word frequencies. To achieve such comparisons for lexicography there is a need for large corpora that cover substantial samples of each significant variety of a language, so that the lexicographer does not miss words or patterns of word use from a variety of genres (Biber, 1990: 263).

However, what such varieties are and in what proportion they have to appear in a corpus is usually not clear. Central to our argument in this book is that the capturing of different varieties in a corpus can be determined quantitatively and qualitatively. Therefore statistical approaches of judging how good different corpus collection strategies are at providing good coverage are used. The methodology we adopt has been characterised by Leon (2005: 36/37) borrowing from Leech (1991: 106/107) thus:

- Focus on linguistic performance, rather than competence;
- Focus on linguistic description rather than linguistic universals;
- Focus on quantitative, as well as qualitative models of language;
- Focus on a more empiricist, rather than a rationalist view of scientific inquiry.

To carry out experiments we need the following:

i. First, one needs a language to work with. For this study we have selected the Setswana language.

ii. Second, one needs a corpus of such a language comprising samples of different text types on which experiments can be performed. For experimentation, a 13 million-word Setswana corpus with a variety of text types on which experiments will be carried out has been compiled. The intended purpose of the corpus is defined as the aiding of Setswana dictionary compilation and research. While the corpus may be used for

other kinds of linguistic research such as language variation and general linguistics, the corpus is primarily constructed for lexicographic purposes. Narrowing the purpose of the corpus to dictionary compilation and research is significant since it has implications on the kind of mark-up that needs to be undertaken on the corpus and the variety of text types that have to be included in the corpus design. The sampling of the intended corpus has been inspired by that of the BNC (Burnard, 1995 and Crowdy, 1991). The aim is to compile a synchronic corpus with texts from 1966 (post-Botswana independence). However, since texts covering broad varieties in Setswana are few, all texts have been considered for inclusion. The scarcity of texts in many categories, e.g. non-existence of newspapers[1], magazines, journals, and other printed matter in many African languages appears to have been a source of discouragement for tackling corpus design in many languages. The corpus that we have compiled is general, in that it is not restricted to any particular subject field, register or genre. Since its use is to test language for general language dictionaries, the corpus comprises a variety of text types from both spoken and written language.

iii. Third, one needs ways of determining how good a corpus is for lexicography. For this research we use keyword analysis, frequency lists and the measure of word types at 10,000 tokens intervals.

The statistical analysis is conducted by the use of a corpus querying software; WordSmith Tools (Scott, 2004-2006) which is an integrated suite of three main programs: wordlist, *Concord* and *Keywords*. The wordlist tool can be used to produce wordlists or word-cluster lists from a text and render the results alphabetically or by frequency order. It can also calculate word spread across a variety of texts. The concordancer, *Concord*, can give any word or phrase in context – so that one can study its co-text, i.e. see what other words occur in its vicinity. *KeyWords* calculates words which are key in a text i.e. used much more frequently or much less frequently in a given corpus than expected in terms of a general corpus of the language.

[1] The *Naledi ya Botswana* newspaper which dates to the 1940s and *Mokgosi* newspaper of 2002-2005 have both seized distribution.

In our experiments keywords are first calculated for the different text types. Because of space constraints, the top 100 keywords of the test from each text type are given. The top 100 keywords constitute a limited version of the total results, however they are sufficient to advance and illustrate the line of argument we are pursuing. Second, type token measures of text types are calculated at comparable 10,000 token intervals. The aim is to determine lexical richness of text types at comparable points. The results shed a light on whether text types with a similar number of tokens have different word types. The significance of this experiment is in demonstrating that individual text types alone are limited in generating broad coverage word types which can be used generating a headword list. On the other hand, text types collectively complement each other in the word types they contribute. While certain text types may display a low number of types at 100,000 token intervals, such low types may be specialized and unique to the text type and therefore be valuable to the entire corpus.

Our argument is that for a corpus to represent a language, it must be designed in such a way that it includes a variety of text types from the language which it represents. The inclusion of such varieties of text types should be seen to be balanced. We discuss the subject of corpus balance and representativeness in Chapter 4. We will measure through keyword analysis if and to what extent different text types generate different keywords that are particular to them. The retrieval of unique word types from a text type gives support to the argument that a corpus that captures linguistic variability of a language community must be compiled using a variety of texts drawn from the text types of a language. Representing text variability in a corpus is significant since the quality of corpus-retrieved information for lexicographic purposes depends on the text input at the stage of corpus construction. This position finds support in Dash and Chaudhuri who argue that,

> The decision about what should belong to a corpus and how the selection is to be made virtually controls every aspect of subsequent analysis. If designed methodically, it can reflect the language with all its features and qualities (Dash and Chaudhuri, 2000: 180).

1.5 Aims of the study

The aim of this study is to determine how Setswana corpora should be compiled and structured as balanced and representative entities through both quantitative and qualitative means in order for them to be "better suited" for lexicography. The aim is to measure whether a corpus

compiled with texts from various text types or a corpus compiled with texts from few or a single text type generates words that are equally good for lexicography. We proceed from the assumption that text variability in corpus compilation is desirable. The assumption, however, demands empirical verification. Such verification can be achieved through experimentation which compares corpora and corpora components. To perform such comparisons accurately, we employ statistical methods since we agree with Kilgarriff (2000: 109) that "lexicographers need the skills and or the software to navigate through sometimes huge numbers of corpus instances." They need to apply statistical methods and natural language processing skills to make sense of the data. Such skills have been demonstrated in Bharathi et al. (2002). Bharathi et al., discuss the statistical analysis of ten Indian languages. The analysis is conducted using basic statistics like unigram frequencies, bigrams frequencies, syllable frequencies, word length distribution and sentence length distribution in the corpora of the ten languages. They were able to extract the following from the corpus (i) word frequencies and their percentages in the whole corpus (ii) the number of distinct words required to cover a certain percentage of corpus (iii) syllable frequencies and pattern extraction from syllables (iv) entropy of words in the corpus (v) word length analysis using average word length, modal word length and (vi) sentence length analysis using average sentence length, modal sentence length, etc.

The aim of this study is to determine if different text types contribute distinct word types. If this is found to be the case then such evidence would prove significant to corpus design for lexicography in general. The recognition that different text types contribute different words, would then influence lexicographers, compiling dictionaries on the basis of corpus evidence, to pay particular attention to corpus design to ensure the broadest coverage possible of text types.

1.6 Research goals

In this research it is aimed to develop a model of corpus construction for the Setswana language which will provide a blue print for corpus design for languages similar to Setswana.

It is also the aim of this study to develop a structured Setswana corpus comprising a variety of text types to be used for experiments in this study and for future research of the Setswana language in size and context.

We aim to calculate and extract through keyword analysis words which are typical of different Setswana text types.

We aim to use frequency analysis to analyse and compare Setswana text types. Frequency will also be used to compare the Setswana corpus and the British National Corpus.

We aim to measure and determine whether the representation of linguistic varieties in a corpus is crucial to a corpus output that reflects linguistic variability or whether similar outcomes may be achieved through building an archive of texts from a single genre.

1.7 Exposition of chapters

Following the introductory Chapter 1, the Setswana language is discussed in **Chapter 2**. In Chapter 2 the different contexts in which the Setswana language is used are examined. The different varieties of Setswana are relevant to corpus design, since what is modelled in a representative corpus is a corpus that reflects linguistic variability. We conclude the chapter by taking a historical view of Setswana research in general and of the development of Setswana lexicography.

In **Chapter 3** we explore corpus lexicography by discussing what a corpus is and whether Web text qualifies as corpus material. Corpus applications on macro- and micro-structural levels are also discussed. We also introduce the exploration of corpora through frequency and keyword analysis and concordance lines inspection. The relevance of corpora to lexicography is discussed and we also examine some pre-electronic corpus studies and some early electronic corpus research. We conclude the chapter by reviewing a variety of methods of headword list identification and the previous use of corpora in Setswana dictionary compilation.

Chapter 4 explores a variety of issues in corpus design for lexicography. These are corpus balance and representativeness, corpus annotation, sample size, and spoken language in a corpus. These are followed by a discussion of 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 in the Setswana language dictionaries. We conclude the chapter by reviewing the Brown Corpus and British National Corpus, illustrating their different strengths and weaknesses.

Chapter 5 discusses the Setswana corpus compiled during this study by examining texts included in the corpus components. The subcorpora types, tokens, type/token ratio (TTR) and standardized type/token ratio (STTR) are calculated.

Chapter 6 and **Chapter 7** are experiment chapters. In Chapter 6 we measure the different subcorpora through keyword analysis determining

which words are typical of the various subcorpora. We demonstrate that different subcorpora are characterised by different keywords. In Chapter 7 we measure how for each text type the numbers of word types grow with every additional 10,000 tokens. The experiment is significant in that it measures types in a variety of text types at similar numerical intervals making it possible to make useful comparisons between the text types.

Chapter 8 concludes and summarises the findings of this study.

CHAPTER TWO

THE SETSWANA LANGUAGE

2.1 The Botswana language situation

In this chapter the position of Setswana within a multilingual Botswana is discussed, situating it within a diverse national linguistic culture.

Botswana, a former British protectorate, is a landlocked southern African country. It has a population of about 1.7 million (2001 census)[1] in a land mass over twice the size of the United Kingdom (Botswana is 600, 370sq km while the United Kingdom is 244,820sq km)[2].

Botswana has an estimated 20 different languages spoken within her borders (Anderson & Janson, 1997: 7). Nyati-Ramahobo (1999: 80) estimates at least "22 distinct languages spoken in the country." These include amongst others: Khoisan languages (!Xoo, Nama, Kxoe!, Shua and others) Setswapong, Thimbukushu, Sekgalagadi, Shiyeyi, Otjiherero, Ikalanga, Setswana, English and many others. Despite its multicultural composition, only two languages, Setswana and English, occupy a dominant position in the educational setting (Mooko, 2004: 181/2). English is the official language and a language of considerable prestige, while Setswana, the language of the dominant Tswana peoples, is the national language and a 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, excluded from being used in the media (both broadcast and print, save for Ikalanga which is used minimally in the *Mmegi* newspaper insert, *Naledi*), parliament, and in most public domains to communicate government policy. Minority languages are in general marginalised from any official function. However, in regions where they are the regionally dominant languages, for instance Mbukushu in north-western Botswana, they are usually used in official roles, like communicating with the chief

[1] The Republic of Botswana: Central Statistics Office, http://www.cso.gov.bw/
[2] The Central Intelligency Agency: The World Factbook: www.cia.com

or nurse (Hasselbring et. al., 2001: 32-33). Of the minority languages spoken in Botswana, Ikalanga is the language of the largest minority people. It is spoken mainly in the North-East and Central Districts of Botswana.

Table 1 gives the different language groups in Botswana and their associated ethnic groups together with regions where the majority of speakers are found. There is uncertainty over the exact number of people associated with different languages and dialects in the country. There are very few reliable figures on the sizes of ethnic groups and scholars at best give estimates of sizes of language communities (see Andersson and Janson, 2004, Hasselbring 2000, Hasselbring et. al., 2001). We therefore do not give any specific figures associated with the languages.

Botswana's educational language policy of 1977 is a controversial document which does not recognize and encourage national linguistic diversity. It appears to be based on the belief that linguistic pluralism is a root source of ethnic and national unrest and not that it empowers citizens to meaningfully participate politically, socially and economically. Alidou (2004) has argued that in post-colonial Africa, in avoidance of ethnic wars, African governments ironically retained colonial languages which were viewed as neutral means of communication. She also argues that governments felt that in the interest of national unity, it was crucial that a country rallied behind a single flag, a single constitution and a single local language hence Setswana as a local language was adopted and sponsored by the Botswana government as a national unifying language. As Bagwasi (2003: 213) argues, "[t]he National Commission on Education 1977 states that Setswana is the language of national pride, unity and cultural pride." Alidou (2004) also observes rightly that in former British colonies African languages and English were used transitionally as medium of instruction and English became a dominant language after the fourth grade and the only language in secondary school and higher education. This state characterised by Alidou reflects the Botswana situation where the 1977 language policy entails the use of Setswana as the medium of instruction in standards (i.e. grades) 1 to 4, followed by a change-over to instruction in English from standard 5. A National Commission which reported in 1993 recommended a change in the policy so that English should become the medium of instruction right from the beginning of primary school, thus excluding Setswana from any such role. The government decided that (Republic of Botswana, 1994) instruction in Setswana is to be in the first year of primary education, and thereafter instruction had to be exclusively in English, save in the teaching of the Setswana language.

Table 1: Botswana's linguistic and ethnic structure

Linguistic Category	Language Family Group	Associated Ethnic Groups	Administrative District
SeTswana	Bantu, Southern	Bakgatla	Kgatleng
		Bakwena	Kweneng
		Bangwaketse	Southern: Ngwaketse
		Bangwato	Central
		Barolong	Southern: Barolong
		Batlokwa	South East
		Batawana	North West
		Balete	South East
		Bakhurutshe	Central
IKalanga	Bantu, Eastern	Bakalanga	Kgalagadi
Se-Birwa	Bantu, Southern	Babirwa	Kweneng,
Se-Tswapong	Bantu, Southern	Batswapong	North West
Se-Kgalagadi	Bantu, Southern	Bakgalagadi Bangologa Baboalongwe Bangologa Bashaga Baphaleng	Kgalagadi, Kweneng, North West
Shiyeyi	Bantu, Western?	Bayeyi	North West
Otjiherero	Bantu, Western	Baherero/Banderu	North West
Thimbukushu	Bantu,Western	Hambukushu	North West
Sesubiya	Bantu, Central	Basubiya/ Bekuhane	North West
Nama	Khoesan	Nama	Kgalagadi/Ghanzi
!Xoo	Khoesan, Southern	!Xoo	Kgalagadi & others
Ju/'hoan	Khoesan, Northern	Ju/'hoan	North West
Makaukau	Khoesan, Northern	Makaukau	Ghanzi
Naro	Khoesan, Central	Naro	Ghanzi
/Gwi	Khoesan, Central	/Gwi	Southern/Ghanzi
//Gana	Khoesan Central	//Gana	Central/Ghanzi
Kxoe	Khoesan, Central	Kxoe	North West
Shua	Khoesan, Central	Shua	Central
Tshwa	Khoesan, Central	Tshwa	Central/Kweneng
Afrikaans	Indo-European	Afrikaans	Ghanzi

Source: Selolwane (2004: 5).

2.2 The Setswana language

Setswana is a member of a Sotho subgroup (also referred to as Sotho languages) of closely related Bantu languages found in southern Africa. This group includes Sesotho, spoken in Lesotho and certain parts of South Africa, and Sepedi, also known as Northern Sotho, which is spoken predominantly in the northern parts of Gauteng, around Pretoria in areas such as Polokwane in South Africa. Southern Sotho, Northern Sotho, and Setswana are largely inherently intelligible but have generally been considered separate languages (see also Cole, 1955: xv/xvi).

Setswana has mother-tongue speakers in at least four countries: South Africa, Botswana, Namibia and Zimbabwe. The largest number of speakers is found in South Africa (over 3 million speakers, about 8% of the population) where Setswana is one of the eleven official languages. Zimbabwe has an estimated 29,000 Setswana speakers and Namibia has approximately 6,000. In Botswana, Setswana is spoken by circa one million speakers (70-90% of the population) as a mother tongue (Andersson and Janson, 1997). Selolwane (2004: 4) observes that "…the SeTswana language is the most dominant of all the language groups found in Botswana, with at least 70% of the population identifying it as a mother tongue and another 20% using it as a second language." Seven percent speak other Sotho-Tswana languages (Setswapong and Sebirwa), 9% Ikalanga, 3% Seherero or Sembukushu, 2% Sesarwa (Khoisan), while 1% speaks Sesobea (Chikuhane) and 1% Seyei.

Her observations on the Setswana language are confirmed by Ramsay's (2006) report that 79% of Botswana's population speaks Setswana as a mother tongue. However other data varies considerably. Ramsay's data is from 2001 household census data.

Ramsay's figures were however disputed by Nyathi-Ramahobo (Gaotlhobogwe, 2006) of Reteng[3] in *Mmegi* of Wednesday 10 May 2006. Reteng countered the data with its own estimates. It argued that unrecognized or minority tribes in the country number 1,030,000 or 60% of the total population, while the main tribes number 305,000 or 17.9% of the total population, with the rest (365,863 or 21%) consisting of immigrants. Reteng's data is speculative and cannot be trusted.

Literature on the language situation in Botswana usually makes a distinction between English as an official language and Setswana as a national language in Botswana. Setswana is seen generally as a language of national unity, and English as a language in which government policies

[3] *Reteng* is a Botswana-based minority tribes' non-governmental organization.