UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Corpus design for Setswana lexicography

Thapelo Joseph Otlogetswe

A thesis submitted in accordance with the requirements
for the degree of Ph.D. in African Languages at
The University of Pretoria.
September, 2007.

Promoter: **Prof. D.J. Prinsloo**
Co-Promoter: **Dr. Adam Kilgarriff**

Summary

This PhD thesis 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).

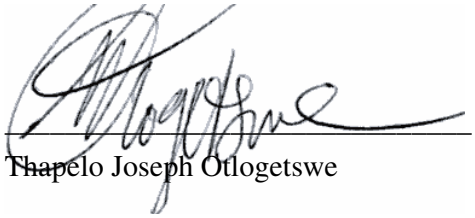
Because of a lack of research in corpora design particularly in African languages, this thesis is an attempt at filling that gap, especially for Setswana. It is hoped that the finding of this study will inspire similar designs in other languages comparable to Setswana.

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.

Declaration

I declare that **Corpus design for Setswana lexicography** is to the best of my knowledge and belief, my original work. All the sources that I have used or quoted have been indicated and acknowledged by means of complete references. The material has not been submitted, either in whole or part, for a degree at this or any other university.



Thapelo Joseph Otlogetswe

Acknowledgements

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 has 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 access to Setswana corpora.

.... Many thanks also to colleagues at ITRI University of Brighton, particularly Ying Ling. ITRI provided me with the finest research atmosphere for my study to flourish earlier in my PhD.

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

… Many thanks to 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.

.... I am thankful to my wife, Shinie Otlogetswe who has been a great source of support during some of the most difficult times of this study.

.... Finally, my unwavering faith in God, through Jesus Christ, has remained a rock and encouragement throughout this study.

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

Table of contents

List of tables

List of figures