

# 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 Prinsloo (2000 and 2001) and Gouws and Prinsloo (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

Summary .....	ii
Declaration .....	iii
Acknowledgements .....	iv
Abbreviations .....	vi
Table of contents .....	vii
List of tables .....	xi
List of figures .....	xiv
<b>Chapter 1 .....</b>	<b>- 1 -</b>
Introduction .....	- 1 -
1.1 Background to the study .....	- 1 -
1.2 Statement of the research problem .....	- 2 -
1.3 Clarifying terms: genre, text type and varieties .....	- 4 -
1.4 Methodology .....	- 6 -
1.5 Aims of the study .....	- 9 -
1.6 Research goals .....	- 10 -
1.7 Exposition of chapters .....	- 10 -
<b>Chapter 2 .....</b>	<b>- 12 -</b>
The Setswana Language .....	- 12 -
2.1 The Botswana language situation .....	- 12 -
2.2 The Setswana language .....	- 15 -
2.3 Setswana dialects .....	- 17 -
2.3.1 The village, cattlepost, lands and city language .....	- 17 -
2.4 Domains of Setswana language use .....	- 18 -
2.4.1 Education .....	- 19 -
2.4.2 Setswana and media .....	- 19 -
2.4.3 The Courts .....	- 20 -
2.4.4 Parliament .....	- 20 -
2.4.5 Churches .....	- 21 -
2.5 Text categories .....	- 21 -
2.6 Challenges of multilingualism and diglossia .....	- 22 -
2.7 The poverty of data .....	- 23 -
2.7.1 The Sanitised Data .....	- 24 -
2.8 Setswana language research .....	- 25 -
2.8.1 A historical overview .....	- 25 -
2.8.2 The development of Setswana lexicography .....	- 26 -
2.8.2.1 Lexicographic tradition .....	- 26 -



2.9 Conclusion .....	- 28 -
<b>Chapter 3 .....</b>	<b>- 29 -</b>
Corpus Lexicography .....	- 29 -
3.1 Introduction.....	- 29 -
3.2 What is a corpus?.....	- 29 -
3.3 Web as corpus .....	- 32 -
3.4 Frequency profiling: frequency and type/token.....	- 36 -
3.4.1 Frequency counts .....	- 36 -
3.4.2 Type/token and word counts.....	- 37 -
3.5 Relevance of corpora to lexicography .....	- 40 -
3.6 Some pre-electronic frequency studies .....	- 47 -
3.7 Electronic-corpora studies .....	- 48 -
3.7.1 An example of frequency profiling.....	- 48 -
3.8 Keyword analysis.....	- 52 -
3.9 Business keywords.....	- 54 -
3.10 Concordance .....	- 56 -
3.11 A review of existing methods of headword list identification.....	- 62 -
3.12 A historical perspective of headword lists .....	- 63 -
3.13 Non-corpus dependant methods of dictionary compilation.....	- 65 -
3.14 Semantic domains .....	- 66 -
3.15 Corpus lexicography and Setswana dictionaries.....	- 68 -
3.16 Conclusion .....	- 69 -
<b>Chapter 4 .....</b>	<b>- 71 -</b>
Issues in corpus design for lexicography .....	- 71 -
4.1 Introduction.....	- 71 -
4.2 Balance and representativeness.....	- 72 -
4.2.1 Proponents of balance and representativeness.....	- 73 -
4.2.2 A cautious approach to balance and representativeness .....	- 77 -
4.3 Corpus annotation .....	- 87 -
4.4 Sample size .....	- 90 -
4.4.1 Spoken versus written corpus text .....	- 93 -
4.4.2 Newspaper text versus the purchase of a pair of shoes.....	- 96 -
4.4.3 The value of spoken language.....	- 98 -
4.4.4 The treatment of borrowings in Toqabaqita.....	- 104 -
4.5 Brown Corpus and BNC review .....	- 112 -
4.5.1 The Brown Corpus .....	- 112 -
4.5.2 The BNC review .....	- 114 -
4.5.2.1 The BNC design criteria .....	- 115 -





4.5.2.2 The BNC written component .....	- 116 -
4.5.2.3 The BNC spoken component .....	- 117 -
4.6 The exploration of both corpora .....	- 119 -
4.7 Conclusion .....	- 120 -
<b>Chapter 5 .....</b>	<b>- 122 -</b>
The Setswana corpus compilation .....	- 122 -
5.1 Introduction.....	- 122 -
5.2 The design strategy .....	- 123 -
5.3 Overall corpus statistics .....	- 124 -
5.4 The Zipfian distribution .....	- 127 -
5.5 Corpus components.....	- 130 -
5.5.1 Text types in the corpus .....	- 131 -
5.5.2 The spoken language components .....	- 132 -
5.5.3 The written language components .....	- 133 -
5.5.4 Newspaper text breakdown.....	- 134 -
5.5.5 Prose text breakdown .....	- 136 -
5.6 The compilation of corpus components.....	- 136 -
5.6.1 Spoken language component compilation .....	- 137 -
i. Sampling .....	- 137 -
ii. Recording.....	- 138 -
iii. Transcription.....	- 140 -
5.6.2 Compiling the written language component.....	- 141 -
i. Sampling .....	- 141 -
5.6.3 Spoken language ethical matters.....	- 144 -
5.6.4 Written language ethical matters .....	- 145 -
5.7 Conclusion .....	- 145 -
<b>Chapter 6 .....</b>	<b>- 147 -</b>
<b>Chapter 6 .....</b>	<b>- 147 -</b>
Measuring text type diversity.....	- 147 -
6.1 Introduction.....	- 147 -
6.2 Keyword analysis.....	- 149 -
6.2.1 Keyword analysis of written components of the Setswana corpus..	- 154 -
6.2.2 Keyword analysis of spoken components of the Setswana corpus..	- 172 -
6.3 Conclusion to keyword analysis .....	- 188 -
<b>Chapter 7 .....</b>	<b>- 191 -</b>
Type/token measures of corpus chunks .....	- 191 -
7.1 Type/token measures .....	- 191 -
7.1.1 The Mean calculation.....	- 194 -



7.1.2 Confidence Interval (CI) calculation .....	- 195 -
7.1.3 Standard deviation .....	- 195 -
7.2 Text divisions for experiments.....	- 198 -
7.2.1 Newspaper Components type/token .....	- 207 -
7.3 Conclusion of type-token measurements .....	- 209 -
7.4 A comparison of the top 100 tokens .....	- 211 -
7.4.1 Comparison of the top 100 tokens of spoken and written Setswana ....	223
7.4.2 Comparison of the top 100 tokens of spoken and written parts of the BNC .....	- 225 -
7.5 A direct comparison of Setswana spoken and written corpus components .....	231 -
7.6 Comparison of opportunistic and balanced corpora .....	- 234 -
7.7 Chapter conclusion.....	- 245 -
<b>Chapter 8 .....</b>	<b>- 248 -</b>
Conclusion and future work.....	- 248 -
8.1 Future research and applications.....	- 256 -
Bibliography .....	- 258 -
Appendix 1: Proposed subentries of <i>pele</i> headword.....	- 276 -
Appendix 2: Participation consent form .....	- 277 -
Appendix 3: Conversation log .....	- 279 -
Appendix 4: Headteacher's letter.....	- 281 -
Appendix 5: Accompanying details for classroom recordings .....	- 284 -
Appendix 6: Letter to publishers asking for text .....	- 286 -
Appendix 7: BNC Part-of-speech codes .....	- 288 -



## List of tables

Table 1: Botswana's linguistic and ethnic structure.....	- 13 -
Table 2: Number of speakers of Botswana languages .....	- 15 -
Table 3: The Setswana text types rendered in the BNC style.....	- 21 -
Table 4: Some of Henry Salt's Setswana terms .....	- 25 -
Table 5: Top 20 words in the Setswana corpus ranked in terms of raw frequency .....	<b>Error! Bookmark not defined.</b>
Table 6: Top 20 words in the Setswana corpus ranked by word spread.....	- 42 -
Table 7: Top 100 Mokgosi sport tokens with functional words.....	<b>Error! Bookmark not defined.</b>
Table 8: Mokgosi sport list's top 100 tokens without functional words .....	- 50 -
Table 9: Mokgosi top 100 sports keywords .....	- 52 -
Table 10: Mokgosi business keywords.....	- 54 -
Table 11: Corpus derived possible subentries of <i>pele</i> entry .....	- 58 -
Table 12: Corpus derived possible subentries of <i>mpa</i> entry .....	- 58 -
Table 13: Corpus derived possible subentries of <i>molomo</i> entry .....	- 58 -
Table 14: Corpus derived possible subentries of <i>lona/dinao</i> entry.....	- 59 -
Table 15: Corpus derived possible subentries of <i>matlho</i> entry.....	- 59 -
Table 16: Setswana days of the week .....	- 107 -
Table 17: Sandiland's rendering of days of the week.....	- 109 -
Table 18: Structure of the Brown Corpus .....	- 112 -
Table 19: The BNC written components .....	- 116 -
Table 20: The BNC spoken components .....	- 117 -
Table 21: Overall corpus statistics.....	- 124 -
Table 22: Top 20 Setswana tokens .....	- 126 -
Table 23: Top 1000 token-ranges and percentages in the whole Setswana corpus .....	- 127 -
Table 24: Top 20 Setswana tokens .....	- 129 -
Table 25: The corpus written and spoken components.....	- 130 -
Table 26: Spoken components statistics .....	- 132 -
Table 27: Overall statistics of the written subcorpus.....	- 133 -
Table 28: STTR measures of the written subcorpus.....	- 134 -
Table 29: Newspaper component statistics.....	- 135 -
Table 30: Prose component statistics.....	- 136 -
Table 31: A contingency table .....	- 151 -
Table 32: Science and technology keywords.....	- 154 -



Table 33: Politics text keywords.....	- 156 -
Table 34: South African Setswana politics terms and Botswana Setswana politics terms.....	- 157 -
Table 35: Poetry text keywords .....	- 158 -
Table 36: Plays text keywords .....	- 159 -
Table 37: Plays text keywords with names treated as metatext.....	- 161 -
Table 38: Grammar texts keywords.....	- 164 -
Table 39: Arts & culture text keywords.....	- 166 -
Table 40: Chat-site text keywords.....	- 167 -
Table 41: News text keywords.....	- 169 -
Table 42: Religious text keywords .....	- 171 -
Table 43: Call-in text keywords.....	- 172 -
Table 44: Face to face dialogue keywords.....	- 174 -
Table 45: Educational spoken text keywords.....	- 176 -
Table 46: Hansard spoken text keywords .....	- 177 -
Table 47: Interviews spoken text keywords.....	- 179 -
Table 48: Open radio programming keywords.....	- 181 -
Table 49: Religious spoken text keywords.....	- 183 -
Table 50: Sport spoken text keywords.....	- 184 -
Table 51: Possible SPORT candidates.....	- 189 -
Table 52: Newspaper types at 10,000 word tokens intervals.....	- 193 -
Table 53: A table of means for Newspaper types .....	- 194 -
Table 54: Newspaper type scores with mean and standard deviation scores .....	- 196 -
Table 55: Newspaper type scores with mean, critical value, standard deviation and confidence interval scores.....	- 197 -
Table 56: Written subcorpus text types .....	- 198 -
Table 57: Three divisions of text types.....	- 199 -
Table 58: Fifteen major corpus text types.....	- 200 -
Table 59: Poetry, Grammar, Chat-site, Plays, POEGRACHAPLA text types .....	- 201 -
Table 60: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types .....	- 203 -
Table 61: Science, Politics, Business, Religious and SCIPOLBUSREL types....	- 205 -
Table 62: Newspaper components types.....	- 207 -
Table 63: Top 100 most frequent tokens in the whole corpus.....	- 212 -
Table 64: Top 100 words: Simple Consistency Analysis results.....	- 214 -
Table 65: Poetry, Grammar, Chat-site, Plays and POEGRACHAPLA.....	- 216 -
Table 66: Science, Politics, Business, Religious and SCIPOLBUSREL .....	- 218 -
Table 67: Prose, Hansard, Call-in, Newspaper and PRONEWHANCAL.....	- 220 -



Table 68: Comparison of written and spoken components to the whole corpus...	223
Table 69: The BNC top 100 words of the whole corpus .....	225 -
Table 70: The BNC top 100 words of the written corpus component.....	226
Table 71: The BNC top 100 words of the context-governed spoken corpus.....	227
Table 72: The BNC top 100 words of the demographic spoken corpus.....	227
Table 73: The BNC top 100 words of the spoken part of the whole corpus.....	228
Table 74: Comparison of the top 100 words of the BNC against the top 100 words of the written and spoken subcorpora.....	229
Table 75: Comparison of BNC and Setswana .....	230 -
Table 76: Outstandingly frequent spoken language.....	231 -
Table 77: Outstandingly infrequent spoken tokens .....	232 -
Table 78: Top 100 tokens of Prose and Combined list.....	238 -
Table 79: Christian terms.....	239 -
Table 80: TSB terms .....	240 -
Table 81: Christian terms and their ranks on the two lists.....	240 -
Table 82: TSB terms and their ranks on the two lists .....	240 -
Table 83: Grammar terms and their position on the two lists.....	241 -
Table 84: Business terms and their rank on the two lists.....	242 -
Table 85: Vulgarities and their position on the two lists .....	244 -



# List of figures

Figure 1: Concordance results of the word <i>pele</i> .....	- 56 -
Figure 2: Word sketch for pray (v) .....	- 89 -
Figure 3: Mantaga concordance lines .....	- 109 -
Figure 4: <i>Sontaga</i> concordance lines .....	- 110 -
Figure 5: A rapid frequency decline in the top 100 words .....	- 129 -
Figure 6: Spoken and written language corpus components pie chart.....	- 131 -
Figure 7: Setswana corpus text types.....	- 131 -
Figure 8: Spoken components statistics.....	- 132 -
Figure 9: Newspaper text division .....	- 135 -
Figure 10: Newspaper types at 10,000 word tokens intervals .....	- 193 -
Figure 11: Prose, Grammar Chat-site, Plays and POEGRACHAPLA types .....	- 202 -
Figure 12: Prose, Newspaper, Hansard, Call-in etc, and PRONEWHANCAL types.....	- 204 -
Figure 13: Science, Politics, Business, Religious and SCIPOLBUSREL types...	- 206 -
Figure 14: Newspaper components types .....	- 209 -
Figure 15: Comparison of the three overall top text types.....	- 210 -
Figure 16: Comparison of the three overall lowest text types .....	- 211 -
Figure 17: 5,000 words from a variety of sources .....	- 237 -
Figure 18: Schematic representation of a balanced corpus construction.....	- 256 -