UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The compilation of corpus-based Setswana dictionaries

Fannie Sebolela

A thesis submitted in accordance with the requirements in the field of study: DLitt: African Languages at The University of Pretoria.

December 2009.

Supervisor: Prof. D.J. Prinsloo

© University of Pretoria

Summary

The aim of this thesis is to describe how corpus-based Setswana dictionaries should be compiled. The challenge to the modern Setswana lexicographer is to compile very practical descriptive and user-friendly dictionaries. A detailed evaluation of existing Setswana dictionaries will be performed in terms of the macrostructural and microstructural aspects:

- Coverage of frequently used words.
- Effective use of dictionary space.
- Use of standard dictionary conventions.
- Choice, ordering and composition of translation equivalent paradigms.

The focus will be on material collection and corpus building. Informants will be used to compile an oral corpus of 100,000 tokens. All ethical requirements such as informed consent requirements (See Appendix 1) will be honoured. Since the text corpus is an organic corpus, thus not a designed corpus aimed at balance and representativeness, the oral corpus will be constructed in the same way i.e. only basic selection criteria:

- Mother tongue speakers of Setswana.

- Adults (to be on a par with authors of the written sources in the text corpus). Age: ranging from 20-60 years.

- Male and female.

Critical analysis of all currently available Setswana dictionaries will be done with special reference to the dictionaries of Brown (1987) (SESD), Snyman, et al. (1990), Matumo (1993).(MSED), Kgasa (1976) (THAND) and Kgasa and Tsonope

(1995).(THAN) In all these cases the strategy would be in terms of the theoretical criteria and best practices in terms of a broad theoretical survey of core aspects of dictionary compilation. Finally, the study will be concluded with an analysis of corpus integrity and stability of Setswana corpora based on the model introduced by Prinsloo and De Schryver (2001a).

Declaration

I declare that **The compilation of corpus-based Setswana dictionaries** is my own work and all sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Fannie Sebolela

Acknowledgements

I am deeply indebted to Professor D.J. Prinsloo for his constructive criticism, purposeful guidance and patience without which this study would probably have been abandoned. May the Lord God preserve you.

My sincere gratitude also goes to Professor P.M. Sebate, Professor M. R. Malope, Dr R. Noormohamed, Ms. R. Ramagoshi, Ms. V. Nicolson and lastly, my dear sister Ms. C. C.L. Ngobeni for correcting and sharing some ideas in this research.

I am also grateful to Mr. M. Phaladi for his willingness to assist in the technical IT support.

Lastly, to my late wife Mmemme and our beloved children, Neo, Motumiseng, Omolemo and my grandson Motheo: You have been there for me from day one. This was a rough one, but finally I have completed the research.

Table of contents

Summary ... ii

Declaration ... iv

Acknowledgements ... v

Table of contents ... vi

List of figures ... x

List of tables ... xii

Chapter 1 Introduction ... 1

1. Background to the study ... 1

1.1 Aims of the study ... 3

1.2 Broad theoretical survey of core aspects of dictionary compilation ... 3

1.3 Critical evaluation of African language dictionaries with specific focus on Setswana dictionaries ... 4

1.4 Designing a comprehensive lexicographic approach for the compilation of Setswana dictionaries ... 4

1.5 Research methodology ... 5

1.6 Scope and limitation of the study ... 6

Chapter 2 Origin and development of the Setswana language ... 8

2.1 Introduction ... 8

2.2 Historical background ... 9

2.3 The role of missionaries towards documentation and the development of the language ... 13

2.4 Population census of the Batswana speaking people ... 17

2.5 Contribution by the missionaries ... 18

2.6 Relationship between Setswana dialects and the standard language ... 19

2.7 Language planning ... 21

2.8 Language change ... 23

2.9 Language development ... 30

2.9.1 Writing system orthography (1929–1937) ... 30

2.9.1.1 Articulator phonetics ... 31

2.9.2 Grammatical rules 34
2.9.2.1 Abbreviation rules 35
2.10 Somerset House Conference of 1947 37
2.10.1 Impact of Bantu education on Setswana orthography 38
2.10.2 Challenges facing the Setswana Language Board 39
2.11 Principles of the Setswana orthography 40
(a) Phonological principle 40
(b) Semantic principle 40
(c) Apostrophe and hyphen 41
(d) Principle to facilitate reading 41
(e) Principle to facilitate writing 41
(f) Principle of the history of orthography 42
2.12 Dictionaries 42
2.12.1 Monolingual Dictionaries 43
2.12.2 Bilingual Dictionaries 44
2.12.3 Trilingual Dictionaries 46
2.13 Conclusion 47
Chapter 3 Compilation of Setswana corpora 48
3.1. Introduction 48
3.2 Major (English) electronic corpora in historical perspective 49
3.2.1 COBUILD main corpus 49
3.2.2 LOB corpus (Lancaster Oslo Bergen) 50
3.2.3 Longman Lancaster English language corpus 51
3.3 Issues ‘size’, ‘balance’, ‘representative’ and ‘organic corpora’ 54
3.3.1 Representativeness and balanced samples 56
3.4 Compilation of the Setswana corpus 56
3.4.1 Text encoding 59
3.4.2 Querying text corpora 61
3.4.3 Alphabetical word list and frequency word list 62
3.5 Compilation of a Setswana oral corpus 64
3.5.1 Keyness definition 65
3.6 Summary of the Setswana oral projects 67

3.7 Identifying and evaluating keyness for the South African oral Setswana corpus79
3.8 Comparison between the South African oral corpus and the South African written corpus 85
3.9 Comparison between the Botswana text corpus and the South African oral corpus 92
3.10 Comparison between the South African written corpus and the Botswana text corpus 98
3.11 Testing the stability of the Setswana corpus 106
3.12 Conclusion 113
Chapter 4 The macrostructure 114
4.1 Introduction 114
4.2 Typical macrostructural inconsistencies existing in Setswana dictionaries 115
4.2.1 Physical limitation on the volume 120
4.2.2 Imbalances regarding the alphabetical stretches 121
4.2.3 Building and applying a multi-dimensional Lexicographic Ruler ... 123
4.2.4 Graphical representation of the macrostructure of the two monolingual Setswana dictionaries 127
4.3 Lemmatisation approaches, strategies and traditions 129
4.3.1 Lemmatization approaches 134
4.3.2 Lemmatization strategies 140
4.3.3 Conjunctiveness versus disjunctiveness 147
4.3.4 Lexicographic Traditions 148
4.4 Problems regarding the lemmatization of nouns with prefixes *le*- and *lo*- in Setswana dictionaries 149
4.5 The absence of nominal derivations 154
4.6 Inconsistency regarding the lemmatization of homonyms 156
4.7 The absence of tonal indication 161
4.8 Dialect words as lemmas in the dictionary are evaluated against the existing Setswana dictionaries 163
4.9 Conclusion 169
Chapter 5 The microstructure 172
5.1 Introduction 172

5.2 Inconsistencies of application within the microstructure of the Setswana dictionaries 174
5.3 Corpora as a key to writing better dictionary articles........ 177
5.3.1 Corpora as an aid to sense distinctions........ 177
5.3.2 Corpora as an aid to retrieve typical collocations........ 180
5.3.2.1 Collocates of *batla* according to the South African Setswana corpus180
5.3.2.2 Collocates of *senka* according to the South African Setswana corpus183
5.3.3 Corpora as an aid to pinpoint clusters 187
5.3.4 Corpora as an aid to select typical and natural examples........ 189
5.4 Problems related to the treatment of the Setswana verbs........ 192
5.5 The treatment of the Setswana months........ 196
5.6 Towards a sound lexicographical treatment of the Setswana months........ 202
5.7 Conclusion........ 208
Chapter 6 Conclusion........ 209
Bibliography........ 214
APPENDIX 1: Summary of the 13 oral Setswana projects........ 229
APPENDIX 2: Pictures of trees in Thanodi ya Setswana........ 230

List of figures

Figure 1: Graphical representation of the South-eastern Bantu Zone 9

Figure 2: Setswana dialects 13

Figure 3: Development of Setswana as a national language 14

Figure 4: Distribution of the Southern Batswana tribes 16

Figure 5: Census in brief: Report No: 03-01-11 (1996) 18

Figure 6: Longman/Lancaster English language corpus – current structure 53

Figure 7: Flow chart representing oral corpus recordings and processing 66

Figure 8: Graphical representation of positive keyness versus frequency in the spoken corpus 82

Figure 9: Graphical representation of negative keyness... 84

Figure 10: Statistical analysis of the Setswana oral corpus in WordSmith Tools......... 85

Figure 11: Statistical analysis of the Setswana written corpus in WordSmith Tools.... 86

Figure 12: Statistical analysis of the Setswana oral corpus in WordSmith Tools......... 92

Figure 13: Statistical analysis of the Botswana text corpus in WordSmith Tools 93

Figure 14: Statistical analysis of the South African text corpus in WordSmith Tools.. 99

Figure 15: Statistical analysis of the Botswana text corpus in the WordSmith Tools 100

Figure 16: A rapid frequency decline in the top 100 words 106

Figure 17: Statistical analysis of the Pretoria Setswana Corpus – Phase 1 (PSetC-Half1) 107

Figure 18: Statistical analysis of the Pretoria Setswana Corpus - Phase 2 (PSetC-Half2) 108

Figure 19: Statistical analysis of the Pretoria Setswana Corpus - Phase 1 (PSetC-Half1) and phase2 (PSetC-Half2) 109

Figure 20: Items when comparing rank ranges of PSetC-Half1 with PSetC-Half2 110

Figure 21: Items when comparing rank ranges of PSetC-Half1 with PSetC-Sum...... 110

Figure 22: Ousted items when comparing rank ranges of PSetC-Half2 with PSetC-Sum 111

Figure 24: A Ruler for Setswana 123

Figure 25: Graphical representation of THAND... 128

Figure 26: Graphical representation of THAN.. 128

List of tables

Table 1: Basic composition of Brown and LOB corpora 51

Table 2: Typical OCR misreading of signs 59

Table 3: Frequency word list reflecting overall counts for the 100 most frequently used words in the South African Setswana text corpus 62

Table 4: Top 100 words from the South African oral Setswana Corpus....................... 77

Table 5: Positive keyness versus frequency in the oral corpus 81

Table 6: Words and their more natural counterparts in the language 83

Table 7: Negative keyness versus frequency in the oral corpus.................................. 83

Table 8: Comparison of the top 100 items between the South African oral and the South African written corpus 87

Table 9: Comparison of the top 100 items between the South African written corpus and the South African oral corpus 89

Table 10: Comparison of the top 100 items between the Botswana text corpus and the South African oral corpus... 93

Table 11: Comparison of the top 100 items between the South African oral corpus and the Botswana text corpora.. 96

Table 12: Comparison of the top 100 items between the South African written corpus and the Botswana text corpus 101

Table 13: Comparison of the top 100 items between the Botswana text corpus and the South African written corpus... 103

Table 14: Summary of SetH1 versus SetH2 112

Table 15: Comparison of the macrostructure between the fixed points rabbit and rally in various dictionaries... 116

Table 16: Overall frequency counts for the derivatives of the verb dira (7053) in WordSmith Tools.. 119

Table 17: A block system for Setswana 124

Table 18: MSED versus the Ruler in terms of page allocation and the number of lemmas... 125

Table 19: Comparison of the alphabetical stretch in THAN and THAND 126

Table 20: Lemmatization approaches, strategies, traditions, etc. 129

Table 21: Complex set of 1-1 relations in the Lemmatization's of nouns and verbs in African languages 130

Table 22: Derivative of reka (buy).. 131

Table 23: Accessibility and information retrieval process for dikagollišano in NSDN137

Table 24: Noun classes and examples of Setswana.. 140

Table 25: Rules for looking up plural forms in Setswana 142

Table 26: Conjunctivism versus disjunctivism... 147

Table 27: Example of data types in a comprehensive monolingual dictionary 175

Table 28: Example of data types in a comprehensive bilingual dictionary 176

Table 29: Concordance lines for thari (210) in the South African Setswana corpus... 177

Table 30: Concordance lines for tshela (1540) in the South African Setswana corpus179

Table 31: Corpus lines for batla (5170) in the South African Setswana corpus.......... 180

Table 32: Collocates of batla generated by WordSmith Tools.................................. 181

Table 33: Top ten collocates of the base batla that collocate immediately to the right of batla in the South African Setswana corpus 182

Table 34: Corpus lines for senka (244) in the South African Setswana corpus 183

Table 35: Collocates of senka generated by the WordSmith Tools............................ 184

Table 36: Top ten collocates of the base senka that occur immediately to the right of senka in the South African Setswana corpus 185

Table 37: Comparison between batla and senka in collocation with gore and tiro 186

Table 38: Comparison of the top ten collocates of batla and senka 186

Table 39: Two-word clusters with batla (451) in the initial position in South African Setswana corpus... 187

Table 40: Three word clusters with batla (451) in the initial position in the South African Setswana corpus 188

Table 41: Collocates of the base tshwanetse (with horizons L5-R5) in the South African Setswana corpus 190

Table 42: Three-word clusters with tshwanetse (909) in the South African Setswana corpus 191

Table 43: Corpus lines for dira (7074) in the South African Setswana corpus........... 195

Table 44: The South African oral Setswana corpus 199

Table 45: Concordance lines taken from the South African written Setswana corpus 201

Table 46: Treatment of the Setswana months.. 202

Table 47: Schematic representation of the Setswana months 207