University of Pretoria etd – Nel J G 2003

Zulu-English Cross-Language Information Retrieval: an analysis of errors

By

Johannes Gerhardus Nel

Submitted in fulfilment of the requirements for the degree of

Magister Informationis Scientia

At the

Department of Information Science
Faculty of Engineering, Built Environment and Information Technology

University of Pretoria, Pretoria

Supervisor: Dr Erica Cosijn
November 2003

©

University of Pretoria etd – Nel J G 2003

Who knows for what we live, struggle, and die? Wise men write many books, in words too hard to understand. But this, the purpose of our lives, the end of our entire struggle, is beyond all human wisdom...

Alan Paton

Abstract

This study is an analysis of the published results of Zulu to English Cross Language Information Retrieval experiments. In these studies, approximate string matching was used to match the running text of the natural language queries to the base forms of the dictionary entries and this aspect was found to be reasonably successful. However, on a conceptual level, Zulu-English CLIR was very poor, in part due to the lack of technical terminology in Zulu. These studies have shown that unique translation problems are to be encountered if the language pairs used contain disparate vocabularies and increases the complexity of CLIR.

In support of the analysis in this study, a detailed overview of the Zulu language as one of the indigenous languages of South Africa is provided, together with a discussion on the linguistic structure of Zulu. Furthermore, the rationale for Zulu-English CLIR is given within the context of the digitisation of indigenous knowledge.

The manual analysis of the query set used in the original experiments shows that two main problems can be identified, namely dictionary problems and translation problems. The dictionary problems are further subdivided into issues relating to orthography, borrowed words and proper names while the translation problems consist of issues relating to paraphrasing, word inflection, homonyms, affixes and mismatched word forms. These two primary problems have further been categorised into problems related either to matching issues, or to culture-related issues.

It is envisioned that interactive CLIR, which will enable a user to select the best-translated keys and add his/her own source keys, might be an ideal solution for Cross-Cultural Information Retrieval. It is certainly worth investigating the following proposed solutions: query expansion, applying metadata to describe the content, applying normalisation, and improving dictionary coverage to manage untranslatable terms.

University of Pretoria etd – Nel J G 2003

I declare that

Zulu-English Cross Language Information Retrieval: an analysis of errors

is my own work, and that all sources applied or quoted have been indicated and acknowledged by means of complete references.

JG Nel

What you put on paper and how you put it there reveal your standards of excellence, your knowledge, and the quality of your thinking more eloquently than anything else about you.

Leedy

What you put on paper and how you put it there reveal your standards of excellence, your knowledge, and the quality of your thinking more eloquently than anything else about you.

Leedy

University of Pretoria etd – Nel J G 2003

Acknowledgements

I would like to express my sincere gratitude to:

- Dr Erica Cosijn, for her invaluable advice, encouragement, motivation, exceptional leadership and insight;
- Prof. Theo Bothma, for the opportunity to study at the Department of Information Science;
- Prof. Hannes Britz (for everything);
- Dr Magdalena Nel, for her motivation and setting an example to follow;
- My parents, for their love, their faith in me, their support and financial assistance;
- My friends, for their prayers and support;
- Clare Hassett, for her wonderful encouragement and faith in me;
- Hilda Kriel, mentor and dear friend;
- Mareli Esterhuysen, for editing and proofreading of the thesis;
- Audrey, Mzangwa, Pansy, Violet and others for their assistance in the translation of the texts; and
- The Lord, for the talent He has given me.

Dedicated to my grandfather
JG Nel, sen.
1929–1997

"Keep your face to the sunshine and you cannot see the shadows."
Helen Keller

University of Pretoria etd – Nel J G 2003

Table of Contents

1 INTRODUCTION 2
1.1 CONTEXTUALISING THE STUDY 2
1.2 PROBLEM STATEMENT 3
1.3 RELEVANCY OF THE PROBLEM TO THE SUBJECT FIELD 4
1.4 RESEARCH PLAN 4
1.4.1 Research methodology 4
1.4.1.1 Literature study 4
1.4.1.2 Qualitative analysis of empirical results 4
1.5 LIMITATIONS OF THE STUDY 5
1.6 TERMINOLOGY 5
1.6.1 Clarification of terms 5
1.6.2 Abbreviations 8
1.7 OUTLINE OF CHAPTERS FOR THE REMAINDER OF THE THESIS 8

2 CROSS-LANGUAGE INFORMATION RETRIEVAL: AN OVERVIEW 11
2.1 INTRODUCTION 11
2.2 PREVIOUS RESEARCH – THE ORIGINS OF CLIR 11
2.3 APPROACHES AND STRATEGIES 13
2.3.1 Dictionary-based methods 14
2.3.1.1 Proper names and spelling variants 15
2.3.1.2 Special terms, technical terms and domain specific terms 16
2.3.1.3 Word inflection 16
2.3.1.4 Phrase translation 18
2.3.1.5 Compound words 19
2.3.1.6 Lexical ambiguity 20
2.3.2 Corpus-based methods 21
2.3.3 Machine translation 23
2.4 CHAPTER SYNOPSIS 25

3 ZULU: A LINGUISTIC OVERVIEW 27
3.1 INTRODUCTION 27
3.1.1 South African languages after 1994 27
3.2 A PROFILE ON THE ZULU LANGUAGE 29
3.2.1 General introduction 29

University of Pretoria etd – Nel J G 2003

3.2.2 Linguistic affiliation 29

3.3 THE LINGUISTIC STRUCTURE OF ZULU 30

3.3.1 The system of noun classes 30

3.3.2 The system of concords 32

3.3.3 Other grammatical structures 32

3.3.4 Term creation in Zulu 33

3.3.4.1 Term creation by means of semantic shift 34

3.3.4.2 Term creation by means of compounding 34

3.3.4.3 Term creation by means of deideophonisation 35

3.3.4.4 Term creation by means of adoptives 35

3.3.4.5 Term creation by means of deverbalisation 36

3.3.4.6 Term creation by means of loan-translation 37

3.4 DICTIONARIES 37

3.5 CHAPTER SYNOPSIS 38

4 RESULTS OF PREVIOUS EMPIRICAL STUDIES 40

4.1 INTRODUCTION 40

4.2 A BRIEF OVERVIEW OF CLEF 41

4.3 METHODS AND DATA USED IN QUERYING THE DATABASE 42

4.4 THE RETRIEVAL SYSTEM AND TEST QUERIES 43

4.5 ZULU TO ENGLISH TRANSLATION 50

4.6 CHAPTER SYNOPSIS 51

5 PROBLEM IDENTIFICATION AND ANALYSIS 54

5.1 INTRODUCTION 54

5.2 AN ANALYSIS OF ERRORS 54

5.3 DICTIONARY PROBLEMS 58

5.3.1 Orthography 58

5.3.2 Borrowed words 60

5.3.3 Proper names 63

5.4 ERROR ANALYSIS OF THE TRANSLATION PROCESS 65

5.4.1 Paraphrasing 65

5.4.2 Word inflection 68

5.4.2.1 Palatalisation 69

5.4.2.2 Pre-nasalisation 70

5.4.2.3 Vowel coalescence 71

5.4.2.4 Vowel elision 71

University of Pretoria etd – Nel J G 2003

5.4.3 Homonyms 72
5.4.3.1 Matching the exact forms of the individual words, as they appear in the running text 77
5.4.3.2 Matching the limited normalised word form 78
5.4.3.2 Matching the normalised inverted index to the dictionary entries 78
5.4.4 Prefixing and suffixing 80
5.4.4.1 Forming of locatives 80
5.4.4.2 Forming of conjunctives 80
5.4.4.3 Verbal extensions 81
5.4.5 Mismatching 82
5.4.5.1 The enclitic 82
5.4.5.2 The interrogative 82
5.5 CONCLUSIONS AND RESEARCH FINDINGS 82
5.6 CHAPTER SYNOPSIS 83

6 TOWARDS CROSS-CULTURAL INFORMATION RETRIEVAL 85
6.1 INTRODUCTION 85
6.2 CATEGORISING THE PROBLEMS IDENTIFIED IN THE ERROR ANALYSIS 85
6.2.1 Issues based on the matching process 88
6.2.2 Issues based on cultural differences 90
6.3 CHAPTER SYNOPSIS 93

7 CONCLUSION 96
7.1 MQ: WHAT ARE THE MAIN PROBLEMS ASSOCIATED WITH THE DICTIONARY APPROACH TO ZULU-ENGLISH CLIR? 96
7.2 SQ1: HOW SUCCESSFUL WERE THESE PARTICULAR EXPERIMENTS IN ZULU-ENGLISH CLIR? 97
7.3 SQ2: WHAT RELIABLE SOLUTIONS CAN BE IMPLEMENTED TO ADDRESS THESE SPECIFIC PROBLEMS? 98
7.3.1 Query expansion 98
7.3.2 Applying metadata to describe the content 100
7.3.3 Applying normalisation 101
7.3.4 Improving dictionary coverage to manage untranslatable terms 102
7.4 SQ3: WHAT CAN BE DONE TO CONTEXTUALISE LANGUAGE AND CULTURE IN TERMS OF CLIR? 104

University of Pretoria etd – Nel J G 2003

7.5 THE ROAD AHEAD 105

8 BIBLIOGRAPHY 108

APPENDICES 118

Appendix A 119

Appendix B 220

Appendix C 225

University of Pretoria etd – Nel J G 2003

We have to do the best we can. This is our sacred human responsibility.

Albert Einstein

University of Pretoria etd – Nel J G 2003

List of figures

Figure 2.1 Query translations via a bilingual dictionary (Pirkola et al., 2001) 14
Figure 2.2 Unrecognized words in morphological analysis (Pirkola et al., 2001) 17
Figure 2.3 Translation ambiguity (Pirkola et al., 2001) 20
Figure 2.4 Parallel corpora in CLIR (Pirkola et al., 2001) 22
Figure 2.5 CLIR based on machine translation (Pirkola et al., 2001) 24
Figure 3.1 Mothertongue speakers in South Africa (Statistics South Africa, 2003) 28
Figure 3.2 The language family tree for Bantu languages 29
Figure 5.1 A comparison between the dictionary problems and translation problems 56
Figure 5.2 A comparison of the types of errors encountered in the translation process 57
Figure 6.1 A schematic representation of the analysis of errors 86
Figure 6.2 A comparison of culture-related issues and matching issues 92

University of Pretoria etd – Nel J G 2003

List of tables

Table 3.1 Number of speakers per language in South Africa (1980, 1991, 1996, 1998 and 2001), (Statistics South Africa, 2003) 28

Table 3.2 Classifying nouns into singular and plural 30

Table 3.3 Different nouns derived from a stem 31

Table 3.4 The different noun classes 31

Table 3.5 Formation of pronouns 32

Table 3.6 Adding verbal extensions (suffixes) to a root give the verb different meanings (Bosch and Taljaard, 1988) 33

Table 3.7 Creating new terms by modifying/expanding the meaning 34

Table 3.8 Creating new terms by combining different words 34

Table 3.9 Creating new terms by adding a prefix to an associated sound 35

Table 3.10 Loanwords in Zulu 36

Table 3.11 Creating new terms by deriving verbs from nouns 36

Table 3.12 Creating new terms by literally translating the word's meaning 37

Table 4.1 A roadmap of CLEF 41

Table 4.2 Matching results for 65 Zulu source words in CLEF Topics C041 to C045 (Cosijn et al., 2002b) 46

Table 4.3 The performance of CLIR queries (Cosijn et al., 2002b) 50

Table 4.4 The effects of untranslated English words on the performance of the best test queries (syn1) (Cosijn et al., 2002b) 50

Table 5.1 The different writing rules related to the sound system of Zulu 59

Table 5.2 Zululised borrowed words 62

Table 5.3 Borrowed words with a class prefix added 63

Table 5.4 Proper names in Zulu 63

Table 5.5 Paraphrasing in Zulu 66

Table 5.6 The effects of palatalisation 69

Table 5.7 The effects of pre-nasalisation 70

Table 5.8 The three principal cases of vowel coalescence 71

Table 5.9 The occurrence of vowel elision in the different classes 72

Table 5.10 Examples of homonyms 73

Table 5.11 Mistranslated words as found in several queries 74

Table 5.12 Comparing the dictionary and mother tongue translation of 'inauguration' 76

Table 5.13 The Zululised amakhemikheli and the (incorrect) resulting matches in the text 76

University of Pretoria etd – Nel J G 2003

Table 5.14 Proposed dictionary matches without a morphological analyser 77
Table 5.15 Matching inflected verbs in different forms 78
Table 5.16 The effect of normalisation 79
Table 5.17 Normalisation and ambiguity 79
Table 5.18 The different verbal extensions occurring in Zulu 81
Table 6.1 Comparing matching problems and cultural problems 87
Table 7.1 Untranslatable words in Zulu 102