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

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.

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

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

Table of Contents

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

Albert Einstein

List of figures

List of tables

1 Introduction

1.1 *Contextualising the study*

When using the term Indigenous Knowledge (IK), people generally refer to traditional knowledge, local knowledge, indigenous technical knowledge, folk knowledge, rural people's knowledge or cultural knowledge. Grenier (1998) defines IK as the unique, traditional and local knowledge existing within, which develops around specific conditions of people indigenous to a particular geographical area. The International Institute of Rural Reconstruction (International Institute Of Rural Reconstruction, 1996) views IK as the knowledge that people in a given community have developed over time, and which they continue to develop. This dynamic type of knowledge is usually based on previous experiences, is often tested over centuries, and adapted to the local culture and environment.

From these definitions, and for the purpose of this study, the following definition is presented:

IK is the *sum total of traditional knowledge and skills* possessed by the people belonging to a particular geographic area, which enables them *to benefit from their natural environment*. IK is *local knowledge* that is *unique to every culture and society*. This knowledge is *embedded* in the decision-making processes, community practices, relationships and rituals of the various tribes and cultural groups found all over the world. Such *dynamic* knowledge and *innovative* skills and practical experiences are shared over generations, and each new generation adds and adapts in response to the changing circumstances.

From the definitions presented, it is clear that IK is not only important, but that it is also a valuable resource. This is especially true for the local people of a community. The importance of IK is realized in the sharing, storing, and managing of IK as a resource so that other people can make use of it and learn from it.

In most tribal communities in South Africa, IK is generally not documented; it is rather shared through traditional oral communication systems, in other words – through word-of-mouth. The fact that IK is mainly unspoken also makes it very difficult to record, transfer, and disseminate this knowledge.

The challenge therefore, is to encourage these tribal communities to share their knowledge and make it available, so that IK be documented. In order to make IK more accessible, it must be captured it in such a way that it can easily be shared and documented. The value of digitizing IK and making it available in the form of electronic databases is that it could be made easily accessible and available to a wider audience. A possible option would be to keep the information in the database in the original language. The value of having the information in the original language reflects in the life experiences, cultures and aspirations that make this solution truly and distinctly South African. In addition, it would also be cheaper to keep the information in the original language, since there will be no need to translate the entire database.

On the other hand, if the databases are compiled in the indigenous languages, this would preclude many users from accessing this information due to their inability to speak the language fluently. For example, an English speaking person may well be able to read Zulu, but may not feel confident in expressing a specific query in Zulu. Such users will likely find cross-language information retrieval (CLIR) particularly useful for languages where they are less confident in their ability to express their information needs effectively. CLIR addresses the situation where the user presents a query in one language (e.g. English) to the database which is in another language (e.g. Zulu), in order to retrieve relevant documents. This presents a number of unique challenges, which is addressed in this study.

1.2 Problem statement

The focus of this study was to do an analysis of the problems experienced during Zulu-English CLIR, as well as expanding on the empirical study of Cosijn et al. (2002a, 2002b, 2002c, 2002d) as described in Chapter 4. This study specifically provided explanations as to why the outcomes were not completely satisfactory (see Chapter 5 for more detail).

The main question were asked as follows:

MQ: What were the main problems associated with the dictionary approach to Zulu-English CLIR?

If the results were poor, this study identified the reasons for the poor performance of Zulu-English CLIR, while providing explanations regarding these reasons as well. In addressing the research problem, the following sub-questions were asked:

SQ1: How successful were these particular experiments in Zulu-English CLIR?

SQ2: What reliable solutions could be implemented to address these specific problems?

SQ3: What could be done to contextualise language and culture in terms of CLIR?

1.3 Relevancy of the problem to the subject field

The research conducted resulted in a valuable contribution to current research projects on CLIR to be applied to indigenous languages. The Zulu language was selected for this study specifically because of the large number of mother-tongue speakers. This study was an expansion on the empirical work done on Zulu-English CLIR, and specifically focused on the reasons why the outcomes were not completely satisfactory. This will be discussed in more detail in Chapters 4 and 5.

1.4 Research plan

1.4.1 Research methodology

The method of investigation that was followed in this research study was an in-depth literature study, as well as a qualitative analysis that evaluated the retrieval performance and accuracy of the empirical results.

1.4.1.1 Literature study

As a first step into the investigation of CLIR techniques for indigenous languages in the electronic environment, a non-empirical literature study of the subject field was conducted. This was important, because it assisted in defining the key concepts as well as providing a framework for the research design.

1.4.1.2 Qualitative analysis of empirical results

The basic strategy for query translation applied in this study was a word-by-word processing of the specific query and then (for each source language word) look up its target language equivalents and place them into the relevant target language query. Results obtained during previous research conducted by Cosijn et al. (2002a, 2002b,

2002c, 2002d) were used for logical evaluation and interpretation, as to address the problems previously identified. It is important to note that no new statistical experiments were done in this study. However, where previous empirical research done by Cosijn, et al. (2002a, 2002b, 2002c, 2002d) was drawn from 50 queries; only 35 of these queries would further be investigated in detail (with an in-depth analysis of the most significant problems experienced). This will be discussed in Chapters 4 and 5.

1.5 *Limitations of the study*

This study only addressed dictionary-based and translation problems experienced during the process of retrieving information from indigenous languages, while specifically focusing on Zulu in the electronic environment. The different CLIR technologies were reviewed for better understanding of the study, but the specific focus was on dictionary-based methods of information retrieval. The study was not intended to provide a detailed description of Zulu grammar and semantics, but rather to explain the shortcomings experienced when morphological analysis is not carried out.

1.6 *Terminology*

1.6.1 Clarification of terms

Approximate string matching: a technique in which words are compared on the basis of their phonetic similarity. Phonetic codes are computed for the strings that are compared, and the strings with similar codes are counted similarly (Gadd, in Pirkola et al., 2001).

CLEF: Cross-Language Evaluation Forum (CLEF) supports global digital library applications by developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages (in both monolingual and cross-language contexts). The database used in this study is a result of test-suites of reusable data that was created for benchmarking purposes. CLEF's aim is to create a community of researchers and developers studying the same problems, and to facilitate future collaborative initiatives between groups with similar interests. The final goal is to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness in the global marketplace, (Cross Language Evaluation Forum, 2003).

"By 'cross-language information retrieval', I mean the retrieval of documents based on explicit queries formulated by a human using natural language when the language

in which the documents are expressed is not the same as the language in which the queries are expressed. It is the ability to issue a query in one language and receive a document in another that distinguishes cross-language information retrieval from monolingual information retrieval" (Oard, 2001).

Dictionary translation method: a trivial method in which each term or phrase in the query is replaced by a list of all of its possible translations, all of which are taken to the final query (Peters and Picchi, 1997).

Homonyms: Homonyms are words that are *spelled* the same, but have different meanings, for example-BEAR (animal) and BEAR (to carry). The **apparent** similarities in these words sometimes cause confusion-particularly to non-native speakers (Glossary, 1998).

"Indigenous knowledge: the sum total of the knowledge and skills which people in a particular geographic area possess, and which enable them to get the most out of their natural environment. Most of this knowledge and these skills have been passed down from earlier generations, but individual men and women in each new generation adapt and add to this body of knowledge in a constant adjustment to changing circumstances and environmental conditions. They in turn pass on the body of knowledge intact to the next generation, in an effort to provide them with survival strategies" (Indigenous Knowledge and Development Monitor, 1998).

An indigenous language refers to arbitrary oral symbols by which a social group interacts, communicates and expresses itself. It enshrines the culture, customs and secrets of the people (Indigenous Language Institute, 2001).

InQuery retrieval system: InQuery is a best-match retrieval system but it also allows retrieval of strict Boolean result sets. All result sets, whether matching Boolean conditions or best match queries, are ranked. InQuery is based on Bayesian inference networks and it supports a wide range of operators (including strict Boolean AND, OR, NOT and proximity operators) as well as various best match operators (Allan et al., 2000). The InQuery (TM) software was developed in part at the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst (http://ciir.cs.umass.edu). InQuery (TM) is a registered trademark of Dataware Technologies, Inc.

Interactive query expansion: Refers to techniques where the user has some interaction with the system in the query expansion process (Selberg, 1997).

Metric similarity measure: "A typical metric similarity measure is a real-valued difference function, d, over character strings, which satisfies the conditions

1) $d(s,s') \geq = 0$
2) $d(s,s') = 0 <> s = s'$
3) $d(s,s') = d(s',s)$
4) $d(s,s') + d(s',s'') \geq = d(s,s'')$

for arbitrary character strings s,s',s" (Berghel, 1987).

Morphological parser/analyser: "A *morphological parser* is a tool for going from the surface ("phonetic") representation to an underlying representation, including breaking the word into its 'morphemes', and undoing any phonological rules that have applied" (Maxwell, 1997).

N-gram matching: in n-gram matching text strings are decomposed into n-grams; that is substrings of length n, which usually consist of adjacent characters of the text string. The degree of similarity between the strings is computed on the basis of the number of similar n-grams, and the total number of unique n-grams in the string (Pirkola et al., 2001).

Non-metric similarity measure: In n-gram matching query keys and index entries are decomposed into n-grams, that is, into the sub-strings of length *n*.

Phrase translations: The translation of a group of words as a whole above the word level, normally for inclusion in a bilingual dictionary or as an aid for translators. This is important because idiomatic phrases, collocations, and technical terms often cannot be translated on a word-by-word basis (Fung and McKeown, 2003).

Query expansion strategies: A query expansion strategy is the process where a search engine adds search terms to a user's weighted search with the intent of improving precision and/or recall. The additional terms may be taken from a thesaurus. For example: a search for "car" may be expanded to: car cars auto autos automobile automobiles. (FOLDOC, 1999).

Relevance Feedback: A powerful technique whereby a user can instruct an information retrieval system to find additional relevant documents by providing relevance information on certain documents or query terms (Selberg, 1997).

Source language: the original language in which the query appears.

Structured query model: A model using Boolean operators to express the relations between search keys (Kekäläinen and Järvelin, 1998).

Target language: the language into which one aims to translate the original query.

Tone: To the linguist (or speech therapist) 'tone' is the quality of sound produced by the voice when words are uttered. In a general sense, 'tone' is the attitude of the

speaker (or writer) as revealed in the choice of vocabulary or the intonation of speech (Glossary, 1998).

Translation ambiguity: a result due to many words having multiple possible translations.

Translingual Information Retrieval: Using queries in one language (such as English) to search for documents in different languages (such as German, Italian and Chinese). (The Information Retrieval Group, 2002).

Translingual Information Retrieval (TLIR): it consists of specifying a query in one language and searching document collections in one or more different languages (Translingual Information Retrieval: a comparative evaluation, 1997).

1.6.2 Abbreviations

CCIR	Cross-Cultural Information Retrieval
CIIR	Center for Intelligent Information Retrieval
CLEF	Cross Language Evaluation Forum
CLIR	Cross-Language Information Retrieval
IK	Indigenous Knowledge
IR	Information Retrieval
LCS	Longest Common Substring
MLIR	Multi-Lingual Information Retrieval
MT	Machine Translation
SA	South Africa
TREC	Text REtrieval Conference
TLIR	Translingual Information Retrieval

1.7 Outline of chapters for the remainder of the thesis

In Chapter 1 this study is contextualised. IK is briefly defined and its value as a resource is discussed. Certain problems encountered concerning accessibility are also described. Brief references to previous empirical work are made, and the dictionary-based and translational problems are identified.

In Chapter 2 the concept of CLIR is investigated and discussed. Some of the aspects to be addressed include: definitions as discussed in the literature; characteristics of CLIR problems; the limitations associated with CLIR; and different techniques, approaches and strategies for CLIR.

In Chapter 3 a detailed overview of the Zulu language as one of the indigenous languages of South Africa will be provided, together with a discussion on the linguistic structure of Zulu. The significance of Zulu dictionaries in the Zulu language will be mentioned briefly.

Chapter 4 contains a review and summary of the empirical work that was done for this study. The key qualitative results and findings of this study are presented and compared to results reached in similar investigations reported in the literature. The research methodology, retrieval system and a critical analysis on the empirical data obtained from previous experiments conducted will also be discussed. This will include an explanation of examples of queries (in both English and Zulu test queries) used in the simulated query runs. Furthermore, the n-gram matching technique will be discussed in more detail.

Chapter 5 contains the failure analysis, where the dictionary-based problems (incompleteness of the dictionary, orthography, borrowed words and proper names) and problems experienced during the translation process (paraphrasing, word inflection and homonyms) will be discussed. Other morphological, semantic and orthographical problems that occurred in the experiments with the Zulu queries will also be discussed as part of the analysis of errors found in previous experiments.

Chapter 6 is dedicated to concept of Cross-Cultural Information Retrieval and a short definition of what it encompasses will be provided. Furthermore, this chapter will also present a concise summary of dictionary and translation problems (identified in Chapter 5) in terms of two new categories (matching issues and culture-related issues) to which they could belong.

In Chapter 7 the findings will be summarized and evaluated against the original problem statement and research objectives. Proposed solutions for future research will be identified and presented to the reader.

Chapter 2

The process of succeeding can be seen as a series of trials in which your vision constantly guides you toward your target while in your actual performance you are regularly slightly off target. Success in any area requires constantly readjusting your behavior as the result of feedback from your experience.

Michael Gelb, Tony Buzan

2 Cross-Language Information Retrieval: an overview

2.1 Introduction

Due to the ongoing development of multilingual information retrieval systems, researchers within the information retrieval community are becoming increasingly more interested in the problem of Cross-Language Information Retrieval (CLIR) (Hull, 1997). The basic idea of CLIR is to cross the language boundary by providing access in one language (the source language) to documents written in another language (the target language) (Hull and Grefenstette, 1996; Oard and Diekema, 1998; Hull, 1997). As Internet resources (such as the world wide web) are increasingly available to more and more countries, researchers find it more appropriate to explore document collections that have not been written in their native language. Hull and Grefenstette (1996) also reason that "technological advances have overcome the network, interface and computer system differences that have obstructed information access in the past."

There are three main ways in which Cross-Language Information Retrieval approaches attempt to "cross the language barrier". Through

- query translation,
- document translation, or
- both (Oard, 1997).

Each of these methods will be described in more detail in Sections 2.3.1–2.3.3. This study (as described in Chapters 4 and 5) focuses on the use of query translation, and proposes various ideas to address some of the problems (such as ambiguities and vocabulary coverage) associated with dictionary-based translations (see Section 2.3.1).

2.2 Previous research – the origins of CLIR

Cross-Language Information Retrieval (CLIR) research started out with experiments using controlled vocabularies, associated dictionaries and thesauri (Pirkola, 2001). Currently, free text approaches dominate CLIR experiments. According to Pirkola et al. (2001) free text methods (see Section 2.3) can be further classified according to the resources used to cross the language boundary. This may be corpus-based resources (Section 2.3.1), or machine translation (Section 2.3.3). A significant

number of Cross- Language Information Retrieval approaches focus on existing linguistic resources and which are mainly machine-readable bilingual dictionaries (Section 2.3.2).

Previous research done by Hull and Grefenstette (1996) proved "that dictionary-based translation, where each term or phrase in the query is replaced by a list of all of its possible translations, represents an acceptable first pass at CLIR." In terms of the framework of controlled vocabulary systems, CLIR has been a subject of study for a long time (Hull, 1997). Earlier research by Salton (as cited in Hull, 1997) demonstrated that CLIR systems were equal in performance to monolingual performance when provided with a carefully, manually constructed bilingual thesaurus. However, Hull (1997) stated that "controlled vocabulary systems are less than equal for modern text retrieval for a number of reasons." For instance, the size and dynamic nature of modern information resources makes manual document indexing and thesaurus construction a difficult task. Oard's survey paper (as cited in Hull, 1997) extensively reviews the history of research on CLIR.

CLIR is considerably more complex than traditional Information Retrieval (IR) because some method for translating the document that needs to be queried or using document-ranking algorithms on must be developed. Several approaches have been proposed and tested, including using resources such as:

- Bilingual dictionaries (Hull and Grefenstette, 1996; Grefenstette, 1998; Ballesteros and Croft, 1996; Ballesteros and Croft, 1998b; Davis, 1998; Davis and Ogden, 1997),
- Thesauri (Mateev et al., 1996; Sheridan et al., 1997; Sheridan and Shäuble, 1997; Sheridan and Ballerini, 1996),
- Corpora (Rehder et al., 1998; Littman et al., 1998; Landauer and Littman, 1991; Carbonell et al., 1997; Peters and Picchi, 1997) and
- Machine translation systems (Yamabana et al., 1998; Gachot et al., 1998; Carbonell et al., 1997)

Some of the more recent theoretical research for using these strategies in terms of CLIR will be reviewed in this chapter.

2.3 *The main approaches and strategies applied in CLIR*

The basic approaches in CLIR involve

- *query translation* from the source language into the target language; and/or
- *document translation* from the target language (Pirkola, 2001).

There are several applications or scenarios in which the user of a retrieval system may be interested in finding information on (written in a language other than the user's native or preferred language). In some applications, a user may want to retrieve all possible relevant information in a multilingual text base, irrespective of the language in which the documents appear. For instance – this may occur when searching certain collections of traditional medicine information. In other cases a user may even have some language comprehension ability regarding the language the documents appear in, but the user may not have a sufficiently, rich active vocabulary in the document languages to confidently specify queries in those languages. In this instance, a cross-language search that permits the user to specify target language queries, but retrieve documents in their original language (the source language) is useful. Even if the user cannot read the retrieved documents, he/she at least has a relevant set of retrieved documents that can be manually translated.

In CLIR the main strategies for query translation are based on three different methods, which are:

- *Dictionary-based methods* with specific relevance to (bilingual) translation dictionaries (Oard and Diekema, 1998; Ballesteros and Croft, 1997 and Davis, 1996 and Hull and Grefenstette, 1996 and Ruiz and Srinivasan, 1998),
- *Corpus-based methods* (Davis, 1996 and Davis and Dunning, 1995 and Ruiz and Srinivasan, 1998), and
- *Machine translation*

A combination of these methods can also be applied (Hull, 1997). These methods will be discussed in more detail below, while also highlighting the problems with each specific method. The specific problems experienced in this particular study will further be analysed in Chapter 5, while solutions will also be proposed in context with the experiments done in this study.

2.3.1 Dictionary-based methods

In dictionary-based approaches, bilingual dictionaries are used for query translation. A very basic strategy for the query translation is to process the queries word-for-word and, for each source language word, look up the target language equivalents and place them into the target language query (see Figure 2.1). Retrieval results improve if the target language is structured by synonym groups (Pirkola, 1998), but these groups must each contain all the translations of a single source language word as given by the dictionary.This will be explained in more detail in Section 3.2.

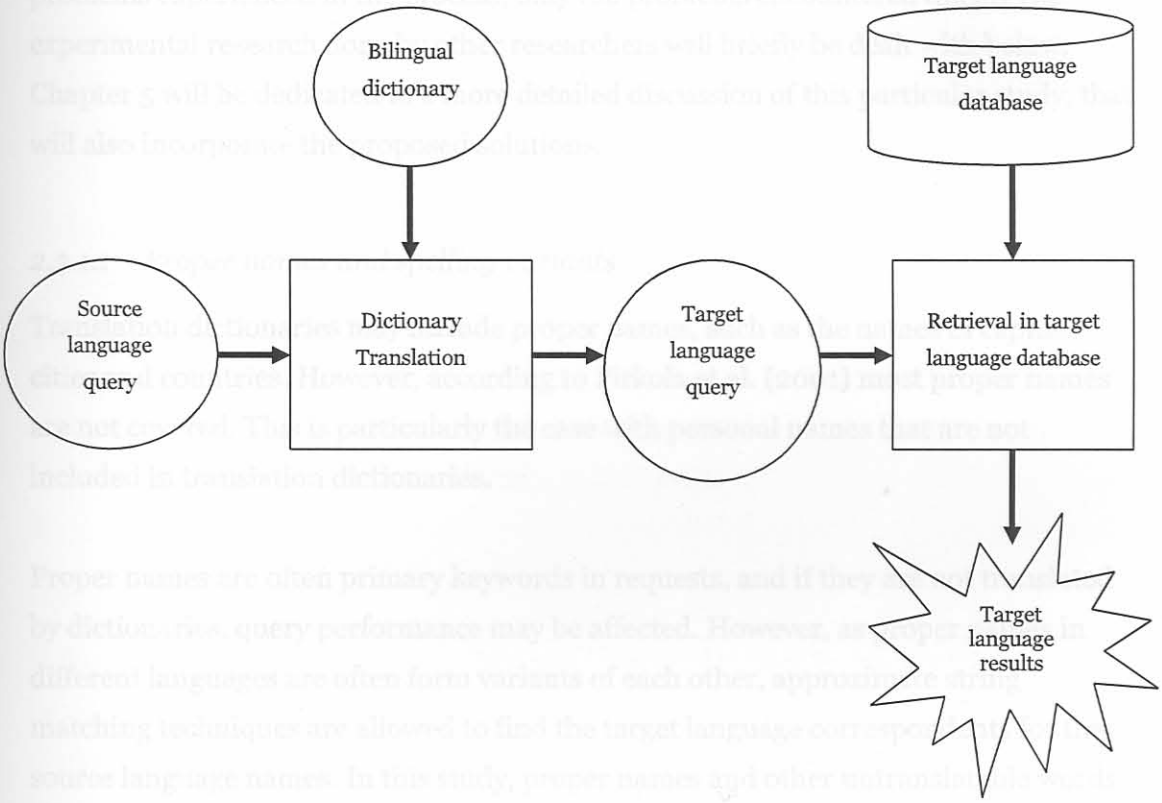

Figure 2.1 *Query translations via a bilingual dictionary* (Pirkola et al., 2001)

Several problems can be associated with the dictionary-based approach to CLIR, and Pirkola et. al (2001) identified the following as some of the most significant obstacles:

- untranslatable search keys (proper names and spelling variants) due to the limitations of translation dictionaries;
- processing of inflected words;
- phrase and compound word identification and translation; and

- lexical ambiguity in source and target languages (Ballesteros and Croft, 1997; Hull and Grefenstette, 1996).

Ballesteros and Croft's dictionary-based work tries to reduced translation ambiguity, by exploring the value of pre- and post-translation query expansion strategies. In doing so, their research supports the findings of Hull and Grefenstette (1996) that phrase translations are important to CLIR.

As the focus of this study is placed on the dictionary-based approach to CLIR and the problems experienced in the process, only the problems encountered during the experimental research done by other researchers will briefly be dealt with below. Chapter 5 will be dedicated to a more detailed discussion of this particular study, that will also incorporate the proposed solutions.

2.3.1.1 Proper names and spelling variants

Translation dictionaries may include proper names, such as the names of capital cities and countries. However, according to Pirkola et al. (2001) most proper names are not covered. This is particularly the case with personal names that are not included in translation dictionaries.

Proper names are often primary keywords in requests, and if they are not translated by dictionaries, query performance may be affected. However, as proper names in different languages are often form variants of each other, approximate string matching techniques are allowed to find the target language correspondents for the source language names. In this study, proper names and other untranslatable words are controlled by an advanced n-gram method (Pirkola, 2001). The n-gram method finds target language spelling variants for proper names in the source language. Proper name translation and matching in CLIR is complicated because proper names may be similarly inflected to common nouns (particularly in Finnish), and may have suffixes (representing different case and number features, as well as other grammatical categories). However, this is not the case with Zulu, because the proper name is not matched at all.

In morphologically complex languages (such as German, Dutch, Korean, Japanese, Arabic and Turkish), proper name searching in CLIR is further complicated by inflection (Pirkola et al., 2001). For example, the name *Gorbachev* is written as

Gorbatshov in Finnish, and it can also take on several inflectional forms, like *Gorbatshoville* (allative, *to Gorbachev*), *Gorbatshovin* (genitive, *Gorbachev's*).

2.3.1.2 Special terms, technical terms and domain specific terms

"General dictionaries often give many equivalents to a source language word, whereas special dictionaries typically give 1–2 equivalents only. The terms of special dictionaries are often unambiguous" (Pirkola et al., 2001). For these reasons, a special dictionary reduces the translation ambiguity problem. Two highly effective methods to address problems such as translation ambiguity and the lack of dictionary coverage, is domain-specific (for example medical, technological, scientific, social and cultural) and general (not limited to a single class or category) terminologies to translate the queries.

Pirkola et al. (2001) argues that "it is highly probable that the special dictionary/general dictionary approach to identifying and translating special terms in query formulation could be extended to other domains of study with great success." This is evident in the fact that many dictionaries can be used in CLIR translation systems. Although each of these dictionaries might have limited content, together they could cover different fields of study and interest.

2.3.1.3 Word inflection

If the source language appears in inflected form they cannot be readily translated, because they do not match dictionary headwords in base forms. A common method (called "stemming") for managing inflected search keys (and derivationally related keys) is to remove affixes from the word forms (Harman, 1991; Porter, 1980). The output is a common root or stem of different forms that may not necessarily be a real word. In lexicon-based morphological analysis, word forms are normalized into base forms that are real words. Morphological analysis also allows for the splitting of compounds into their component words (see section 2.3.1.5). This enables the matching of source language keywords with dictionary headwords (Pirkola et al., 2001). Alternatively, source language keywords and headwords can be conflated into the same form by a stemmer (Davis and Ogden, 1997). One problem related to stemming is where different headwords may be conflated into the same form. According to Hull (1997), the size of a morphological program's lexicon limits the effectiveness of morphological analysis. It is impossible to exhaustively list all the words of a language in the lexicon, as is the case with most translation dictionaries.

This contributes to the problem of untranslatable search keys.

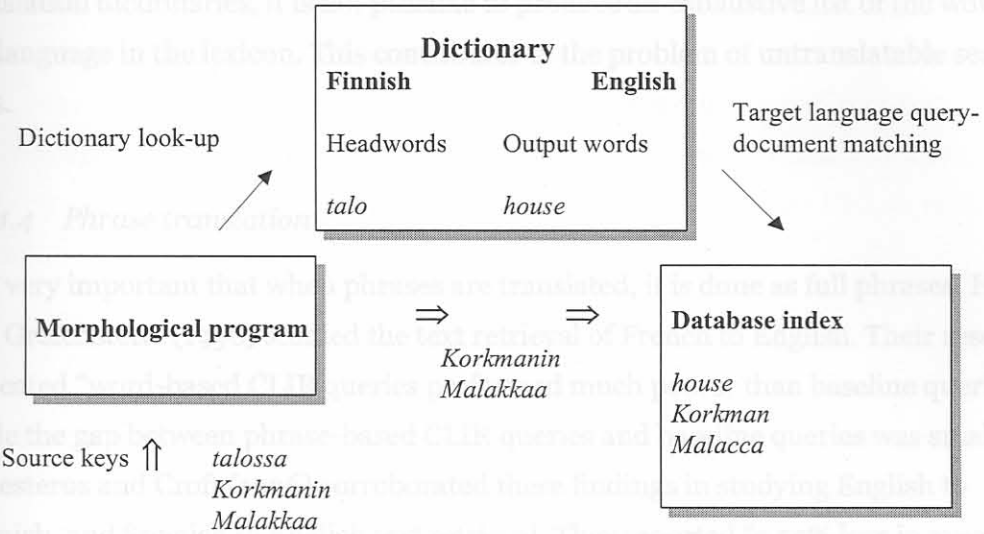

Figure 2.2 *Unrecognized words in morphological analysis* (Pirkola et al., 2001)

In Figure 2.2 Pirkola et al. (2001) portrays how different types of inflected words are managed from dictionary-based Finnish to English CLIR. "Most inflected keys are normalized and translated by a dictionary, like the form *talossa* (in the house) which is first normalized into *talo* and then translated into *house*. The unrecognized forms are sent unchanged into a CLIR query. The names *Korkmanin* (personal name in genitive) and *Malakkaa* (a spelling variant of the geographical name *Malacca* in accusative) represent typical cases of words not listed in the lexicon of a morphological program" (Pirkola et al. 2001). These unlisted words do not match the English index keys. Even if the word 'Malakkaa' were normalized, it would still not match the dictionary entry in the database. Pirkola et al. (2001) mentions "in dictonary-based CLIR, they could be handled similarly as untranslatable spelling variants in the case of a translation dictionary." (See Section 2.3.1.1 for more detail.) In investigating the inflection of words, it was found that CLIR effectiveness also depends on the monolingual component (i.e. the morphological processing of index keys). The effectiveness of stemming obviously depends on the language. Pirkola et al.'s research (2001) generally shows that "*recall* can be expected to improve due to stemming since a larger number of potentially relevant documents are retrieved." The research carried out in different languages (Spanish, Danish, German, French and Portuguese) has shown that stemming improved the *precision* of several of these languages.

Hull (1997) went a step further and proved that the effectiveness of morphological analysis is limited by the size of a morphological program's lexicon. As in the case of translation dictionaries, it is not possible to produce an exhaustive list of the words of the language in the lexicon. This contributes to the problem of untranslatable search keys.

2.3.1.4 Phrase translation

It is very important that when phrases are translated, it is done as full phrases. Hull and Grefenstette (1996) studied the text retrieval of French to English. Their research indicated "word-based CLIR queries performed much poorer than baseline queries, while the gap between phrase-based CLIR queries and baseline queries was small." Ballesteros and Croft (1996) corroborated these findings in studying English to Spanish, and Spanish to English text retrieval. They reported "a 55% loss in average precision for queries translated word-by-word compared with the original queries. A 30% loss in performance resulted from translation ambiguity and a 20% loss was due to inaccurate translation of phrases." In another study, Ballesteros and Croft (1997) showed that the correctness of the translations is just as important in automatic phrase translation. However, if phrase translation fails, phrase-based queries may perform slightly poorer than word-based queries.

"Phrases are not a major problem for languages in which multi-word expressions are compound words rather than phrases, such as German, Swedish, Finnish, and Dutch" (Pirkola 1999). However, if phrases are not identified and translated correctly, the effects on certain queries may prove to be decisive. Ballesteros and Croft (1998b) reported that retrieval performance improves when the phrases are translated by a dictionary. However, not all phrases are listed in dictionaries, which suggests the use of some additional or alternative translation method. Research by Ballesteros and Croft (1998b) and Fujii and Ishikawa (2001) showed that phrase translation, based on word collocation statistics in the target language to be useful. Fujii and Ishikawa (2001) further explored phrase translation in Japanese to English retrieval. In Japanese, as is the case with Zulu, technical terms are often phrases. New technical phrases are generated from existing words (in Zulu this is called 'Zululizing' the term), and the new phrases are not generally listed in dictionaries.

The importance of phrase translation cannot be emphasized enough in CLIR (Pirkola et al., 2001). It is not a problem for languages in which multi-word expressions are compound words rather than phrases, such as German, Swedish, Finnish and Dutch

(Pirkola, 1999). Nevertheless, if phrases are not identified and translated correctly, the effects on certain queries will be negatively impacted.
The results of Hull and Grefenstette's research (1996) done on French-English text retrieval are confirmed by Pirkola et al.'s research findings (2001). The word-based CLIR queries performed much poorer than the baseline queries, while the gap between phrase-based CLIR queries and baseline queries were small. For a review on other recent methods and results of phrase processing in CLIR, see Pirkola et al. (2001).

2.3.1.5 Compound words

Pirkola et al. (2001) defines a compound as "a word formed from two or more words that are written together." In CLIR, the distinction between compositional, non-compositional, semi-compositional phrases and compounds are important. Compositional compounds are "those compounds whose meaning can be derived from the meanings of component words" (Akmajian et al., 1990 as cited in Pirkola et al., 2001).

For instance, "the meaning of the Finnish word *kaupunginhallitus* (*city government*) comes from the meanings of the components *kaupungin* (*city*, in genitive) and *hallitus* (*government*)" Pirkola et al. (2001). As Pirkola phrased it: "in compositional compounds, a full compound typically is a hyponym (a narrower term) of its headword." This is very much the same for the Zulu language as illustrated in the following example: the meaning of the Zulu word *umakhalekhukhwini* (cellular phone) comes from the meanings of the components *u-ma-khala* (cry/ring) and *ekhukhwini* (in the pocket). Pirkola et al. (2001) refers to a compound whose meaning cannot be deduced based on its components, which is a non-compositional compound. The term semi-compositional compound refers to "a compound whose meaning is in part interpretable on the basis of the components" (Pirkola et. al. 2001).

Due to the productive nature of natural languages, words can be combined into any number of new compound words. Some languages such as German, Swedish, Finnish, Dutch and Afrikaans are characterized by a high frequency of compounds (Sheridan and Ballerini, 1996; Hedlund et al. 2001; Pirkola, 1999). For such languages, effective dictionary look-up and the searching for compound words in CLIR cannot only be solely based on full compounds, but must also be based on their component words. The decomposition of compounds and separate translation of

component words are often useful, because translation dictionaries may not include full compounds as such. According to Pirkola (1999), "the separate translation of compositional compounds will give correct senses."

2.3.1.6 *Lexical ambiguity*

It refers to words with the same spelling, but different meanings (homonymy) and words with many different meanings (polysemy). According to Pirkola et al. (2001) "the senses of homonyms are unconnected. A lexeme that has more than one sense is *polysemous*. The word *board*, for example, has several (sub)senses, e.g., (a) *a thin plank*, (b) *a tablet*, (c) *a table*, and (d) *food served at the table*." This example shows that the senses of a polysemous word are related to each other where one sub sense may be a metaphorical extension of another sub sense. Based on morphology, lexical ambiguity can be further divided into base form and inflectional ambiguity (Pirkola et al., 2001). First, *base form ambiguity* describes the condition where two (or more) lexemes–usually two separate headwords of a dictionary–have the same (base) form, as well as describing the condition in which one lexeme has two or more senses. *Inflectional ambiguity* again refers to a condition where two or more lexemes share at least one common inflectional form.

In CLIR, translation ambiguity is primarily caused by lexical ambiguity, which appears as an increase of irrelevant search key senses due to source and target lexical ambiguity. Ballesteros and Croft (1996), Grefenstette (1998), Hull and Grefenstette (1996) and Pirkola et al. (2001) attribute translation ambiguity and difficulty in managing phrases as the main reasons for the low effectiveness of plain dictionary-based CLIR queries.

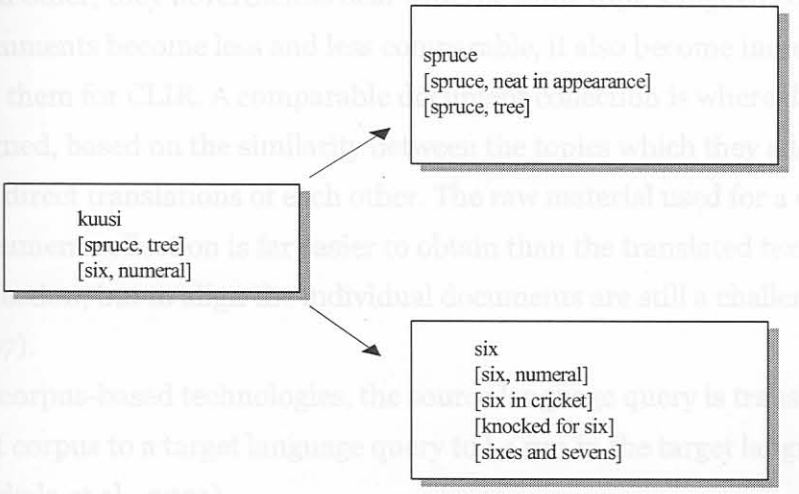

Figure 2.3 *Translation ambiguity* (Pirkola et al., 2001)

Figure 2.3 illustrates the increase of ambiguity in a translation process, as researched by Pirkola et al. (2001).

> The Finnish form *kuusi* is homonymous and has two senses: [spruce, tree], [six, numeral]. The English word *spruce* has two and the word *six* four senses (Collins dictionary, 1998): [spruce, neat in appearance], [spruce, tree], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens]. Think that in monolingual Finnish and monolingual English searching the correct sense is [spruce, tree]. There is one extraneous sense in Finnish, [six, numeral], as well as English, [spruce, neat in appearance]. But in Finnish to English retrieval there are five extraneous senses, [spruce, neat in appearance], [six, numeral], [six in cricket], [knocked for six], [sixes and sevens].

From this example it can be deduced that lexical ambiguity (associated with CLIR queries) stems in part from a source language and in part from a target language.

2.3.2 Corpus-based methods

Corpus-based systems use parallel and/or comparable corpora for query translation (Hull, 1997). This method is also independent of dictionaries. A parallel corpus consists of pairs of documents, where the one document is in the source language of the user query and the other in the target language. Furthermore, the document pairs are translations of each other. Although the documents are not exact translations of each other, they nevertheless deal with the same topic (aligned corpus). Yet, as documents become less and less comparable, it also become increasingly difficult to use them for CLIR. A comparable document collection is where documents are aligned, based on the similarity between the topics which they address because they are direct translations of each other. The raw material used for a comparable document collection is far easier to obtain than the translated text used in parallel collection, but to align the individual documents are still a challenging task (Oard, 1997).

In corpus-based technologies, the source language query is translated in a parallel text corpus to a target language query to be run in the target language database (Pirkola et al., 2001).

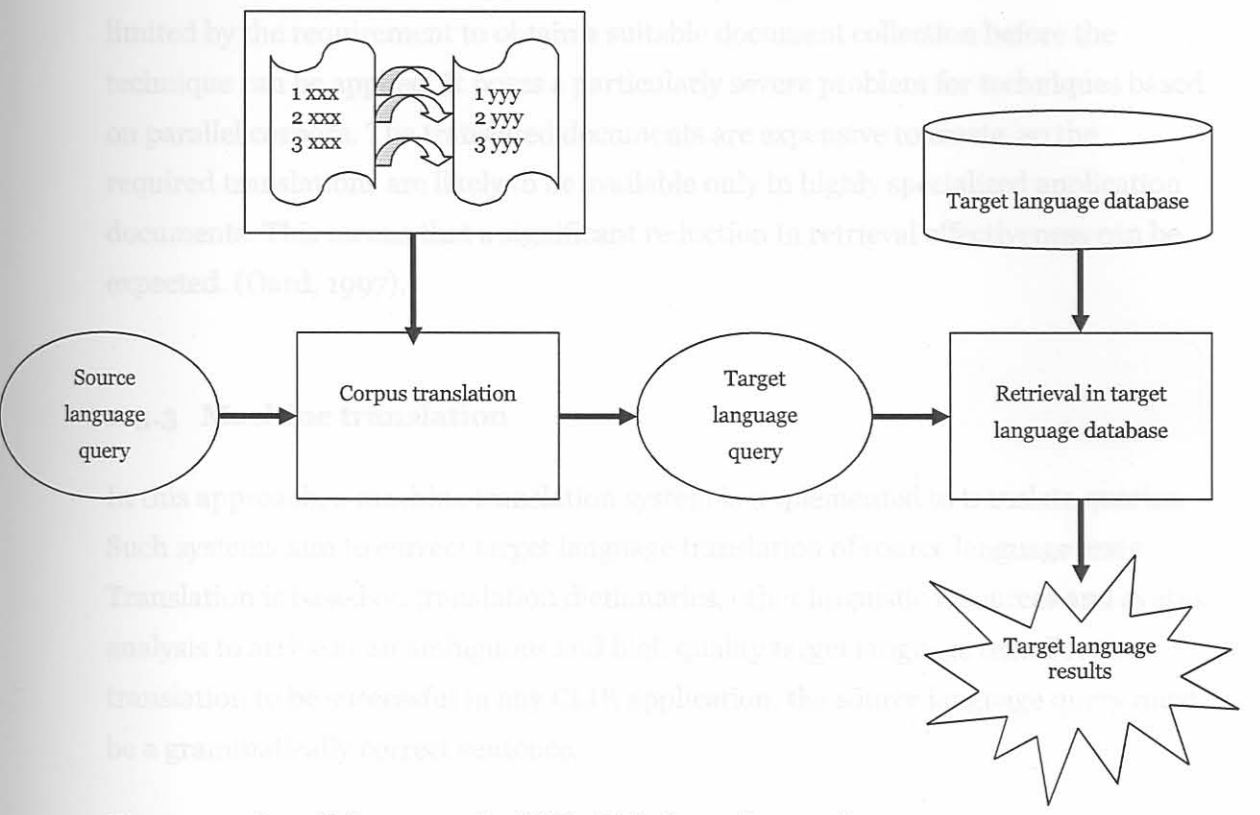

Figure 2.4 *Parallel corpora in CLIR* (Pirkola et al., 2001)

When a source language query is entered into the system, it is run against the source language documents of the parallel corpus. Best-matching documents are identified and their target language pairs retrieved. Statistical criteria and possible sentence-by-sentence alignment are used to identify best topic words to be used in the target language query. The target language query is a bag-of-words query, and run against the target language collection (Pirkola et al., 2001).

Although this approach delivers good results, it has several complications. First, the translation relationship established between parallel words in the text is usually domain dependent (for instance medical, technological, scientific, social, cultural or educational). This means "accuracy decreases outside the domain" (Davis, 1998). Second, because electronic parallel corpora is not readily available in different languages, this approach to CLIR is not practical in a South African context. It is simply too costly to acquire, because it is difficult to obtain existing translations of the correct document type. Furthermore, through the recall performance is high, the precision level is not acceptable. Because of this, the general application of this method should be demonstrated for it to become feasible. Currently, this method also remains very application dependent (Peters and Picchi, 1997).

Furthermore, corpus-based approaches that exploit parallel document collections are limited by the requirement to obtain a suitable document collection before the technique can be applied. It poses a particularly severe problem for techniques based on parallel corpora. The translated documents are expensive to create, so the required translations are likely to be available only in highly specialized application documents. This means that a significant reduction in retrieval effectiveness can be expected. (Oard, 1997).

2.3.3 Machine translation

In this approach, a machine translation system is implemented to translate queries. Such systems aim to correct target language translation of source language texts. Translation is based on translation dictionaries, other linguistic resources and syntax analysis to arrive at an ambiguous and high quality target language text. For the translation to be successful in any CLIR application, the source language query must be a grammatically correct sentence.

Arnold et al. (2001) identified three types of problems encountered in Machine Translation:

- *Ambiguity*: A word is lexically ambiguous when it has more than one meaning. An example of a lexically ambiguous word is "bank", which can refer to both a river bank and a financial institution.
- *Lexical and structural mismatches*: When two languages categorize the same word differently, a lexical or structural mismatch can occur. It is also the case when a concept expressed by multiple words in one language is expressed by one word in another language, or when a word in one language has no equivalent word or lexical unit in another language.
- *Multiword units (idioms)*: Idioms are generally phrases recognized as a unit. These phrases normally have a meaning that differs from the literal meaning of its parts taken together. An example in English would be the phrase "kick the bucket". It is common knowledge that this phrase has the same meaning as the word "die", but this meaning cannot be derived by taking the meaning of each word separately – the phrase must be considered as a whole.

Apart from the mentioned problems generally experienced with this approach, machine translation (MT) systems seem to be a straightforward choice for query translation. For each query *q*, an MT system will give a unique translation *q'* for it. In some instances, the translation is reasonable. But in other instances, the translation may depart from the original query. However, the machine translation system is not highly suitable for CLIR to use as a tool for query translation–as done in this study– for several reasons.

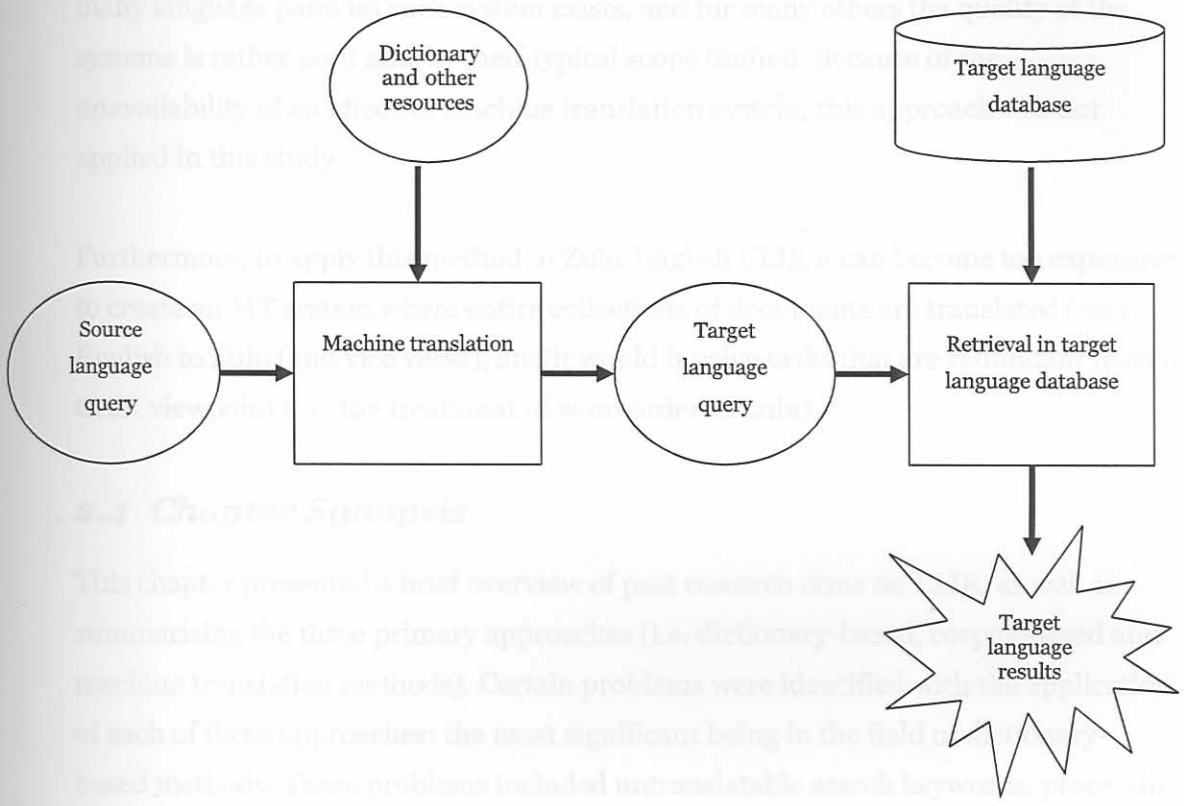

Figure 2.5 *CLIR based on machine translation* (Pirkola et al., 2001)

First, these systems put in a lot of effort trying to produce syntactically and semantically correct senses. Yet, this has no effect on current CLIR approaches, which operate on strings and not meanings (Hasnah and Evans, 1999).

Second, the machine translation system selects one translation of the word. This selection process is not only difficult, but also results in selecting the wrong target translation. In Hasnah and Evans' (1999) opinion, limiting the translation to a single specific word may result in losing several relevant documents that might contain other translations.

Also, typical queries in current IR systems tend to be very short. Therefore, the advantage of MT systems (which in principle can exploit syntactic and semantic aspects of context to improve translation) in respect to dictionary-based approaches is questionable.

Finally, another problem of this approach with regard to the South African context, is the unavailability of good machine translation systems for desired language pairs. For many language pairs no such system exists, and for many others the quality of the systems is rather poor and/or their typical scope limited. Because of the unavailability of an effective machine translation system, this approach was not applied in this study.

Furthermore, to apply this method in Zulu-English CLIR it can become too expensive to create an MT system where entire collections of documents are translated from English to Zulu (and vice versa), and it would involve tasks that are redundant from a CLIR viewpoint (i.e. the treatment of word order in Zulu).

2.4 Chapter Synopsis

This chapter presented a brief overview of past research done on CLIR, as well as summarising the three primary approaches (i.e. dictionary-based, corpus-based and machine translation methods). Certain problems were identified with the application of each of these approaches; the most significant being in the field of dictionary-based methods. These problems included untranslatable search keywords, processing inflected word-forms, identifying phrases and compound words (and translating them), and the lexical ambiguity that occurs in source and target languages. In explaining the different strategies offered by the dictionary-based approach to CLIR, it was concluded that this method would provide the best solution to the identified problems.

The next chapter will focus on Zulu as an indigenous language. The chapter will present to the reader a profile on the Zulu language, with a brief overview of the linguistic structure made up from the different noun classes and concords. The most important part of the chapter will focus on term creation in Zulu, which would help to form a better understanding of the empirical work and results obtained in Chapter 5.

Chapter 3

Aim for success, not perfection. Never give up your right to be wrong, because then you will lose the ability to learn new things and move forward with your life. Remember that fear always lurks behind perfectionism. Confronting your fears and allowing yourself the right to be human can, paradoxically, make yourself a happier and more productive person.

Dr. David M. Burns

3 Zulu: a linguistic overview

3.1 *Introduction*

The topic concerning South African languages can be viewed from different angles. An attempt will be made to establish how the position and use of South African languages, especially Zulu, form part of a broader discourse regarding language, culture and power. However, the primary purpose of this chapter is to create an understanding of the rules of the Zulu language; to analyse the problems and also propose solutions to certain errors experienced during the empirical study described in Chapter 4. The main focus of this chapter is on term creation in Zulu, by:

- means of semantic shifts,
- compounding different words,
- using deideophones or adoptives,
- deverbalisation, and
- loan translation.

By identifying orthographical rules in the Zulu texts, an accurate analysis of errors (in Chapter 5) can be produced.

3.1.1 South African languages after 1994

It is evident that the South African language dispensation has changed quite radically with democratisation in 1994. Before democratisation, languages other than Afrikaans and English did not have the same opportunities for growth and development. This in spite of the fact that approximately 70% of all South Africans have an indigenous language as their mother tongue, whereas 25% have English or Afrikaans as their mother tongue (see Figure 3.1) (Kaschula and De Vries, 2000; Wasserman, 2000).

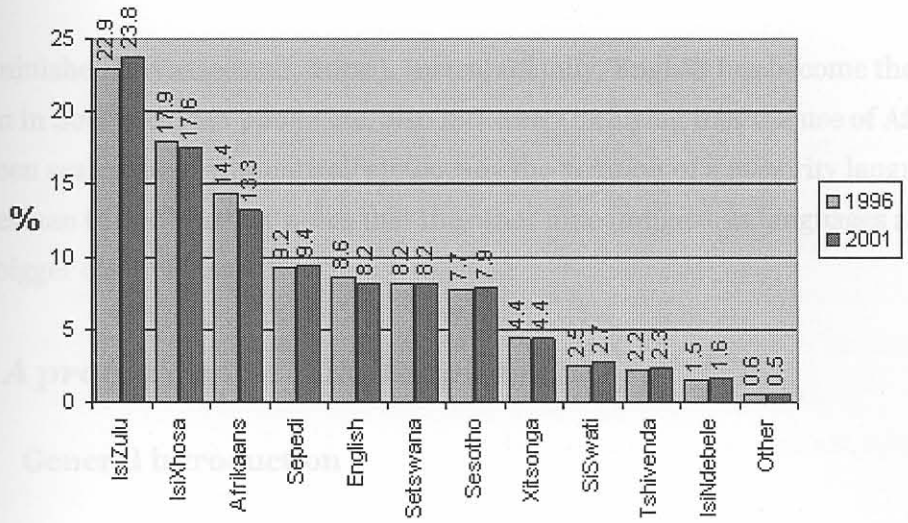

Figure 3.1 *Mothertongue speakers in South Africa* (Statistics South Africa, 2003)

The following table confirms the numbers quoted in Figure 3.1, by stating the figures for each of the language groups since 1990 (while also showing the growth of each language).

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

LANGUAGE	1980	1991 (a)	1996	1998 (b)	2001 (c)
isiZulu	6 064 480	8 343 587	9 200 144	10 194 787	10 677 305
isiXhosa	2 879 360	6 729 281	7 196 118	7 610 435	7 907 153
Afrikaans	4 925 760	5 685 403	5 811 547	5 945 805	5 983 426
Sepedi	2 431 760	n/a	3 695 846	3 832 645	4 208 980
English	2 815 640	3 422 503	3 457 467	3 692 157	3 673 203
Setswana	1 444 908	3 368 544	3 301 774	3 613 925	3 677 016
Sesotho	1 877 840	n/a	3 104 197	3 539 261	3 555 186
Xitsonga	888 140	1 439 809	1 756 105	1 776 505	1 992 207
siSwati	650 600	952 478	1 013 193	1 068 733	1 194 430
Tshivenda	169 740	673 538	876 409	1 227 824	1 021 757
isiNdebele	459 880	n/a	586 961	654 304	711 821
Other	292 360	640 277	228 275	157 767	217 293
Unspecified	n/a	n/a	355 538	10 868	n/a
TOTAL:	26 271 060	31 255 420	40 583 574	43 325 017	44 819 778

The new constitution awarded equal official status (on a national level) to all eleven languages in South Africa. They are: IsiZulu, IsiXhosa, IsiNdebele, Siswati, Sesotho, Setswana, Tshivenda, Afrikaans, English, Xitsonga and Northern Sotho (Sepedi). While this step indicated that indigenous languages are to be developed and their limited rights to be extended, it also stipulated that "existing language rights may not

be diminished" (Wasserman, 2000). Instrumentally, English has become the *lingua franca* in South African public life, also indirectly implying that the use of Afrikaans has been scaled down dramatically to occupy the position of a minority language. Wasserman (2000) further notes that the other nine indigenous languages are at an even bigger disadvantage.

3.2 *A profile on the Zulu language*

3.2.1 General introduction

More than 10 million people in South Africa has Zulu as their home language (UCLA Language Materials Project, 2001; South African Languages–isiZulu, 2001), with the main concentration of people in Kwazulu-Natal. It is also a dominant language in at least a dozen districts in Gauteng and the Free State. Of all the languages spoken in South Africa, including Afrikaans and English, it has the largest number of speakers–comprising of more than 23% of the total population (ETHNOLOGUE report for language code: zuu, 2001).

3.2.2 Linguistic affiliation

Zulu is a Nguni language that includes Xhosa, Swati and Ndebele. These languages are closely related and mutually intelligible. Nonetheless, they are easily considered dialects of the same language for cultural, historical and political reasons. For instance, Zulu and Xhosa have their own identities according to the individual speakers of the respective languages. The Nguni languages are part of a much larger related group of southern Bantu languages, as reflected in Figure 3.2:

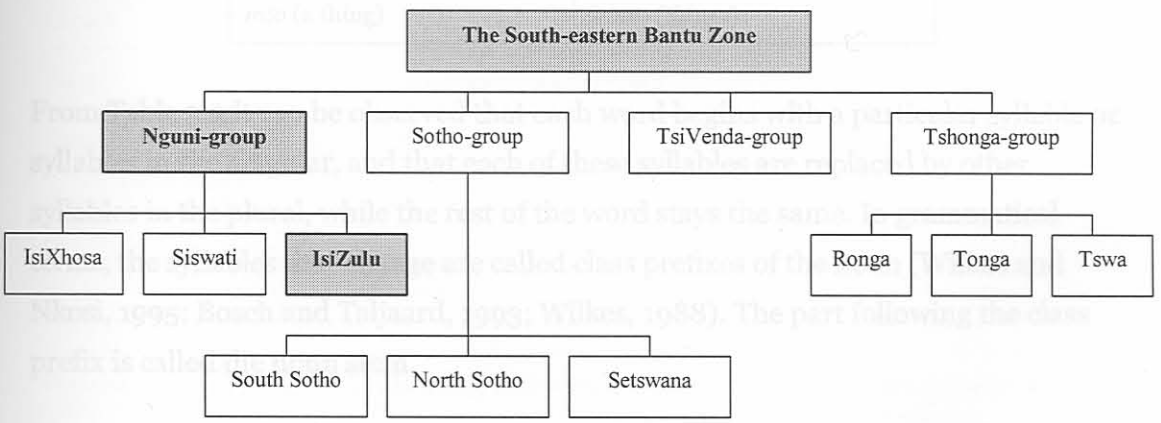

Figure 3.2 The language family tree for Bantu languages

3.3 *The linguistic structure of Zulu*

As Zulu is classified as an agglutinative language, prefixes and suffixes are attached to roots and stems of words, thus conveying the grammatical information. Zulu nouns are divided into sets of classes, called grammatical genders. Each gender has two distinct prefixes, one creating singular nouns and the other plural nouns. The numerous classes far exceed the familiar European classifications (of masculine, feminine and neutral), and are roughly associated with certain semantic characteristics relating to human beings, kinship terms, animals, plants, artefacts, abstract concepts and so on (Cosijn et al., 2002c).

The structure of Zulu is based on two principles (Bosch and Taljaard, 1993; Doke, 1968; Wilkes and Nkosi, 1995): *viz* (a) the system of noun classes, and (b) the system of concords.

3.3.1 The system of noun classes

In Zulu each person or thing, concrete or abstract, is placed in a specific category or group. In grammatical terms, we speak of nouns placed in classes.
Table 3.2 for instance, indicates how nouns are classified into singular and plural forms.

Table 3.2 *Classifying nouns into singular and plural*

Singular	Plural
umuntu (a person)	*abantu* (people)
into (a thing)	*izinto* (things)

From Table 3.2 it can be observed that each word begins with a particular syllable or syllables in the singular, and that each of these syllables are replaced by other syllables in the plural, while the rest of the word stays the same. In grammatical terms, the syllables that change are called class prefixes of the noun (Wilkes and Nkosi, 1995; Bosch and Taljaard, 1993; Wilkes, 1988). The part following the class prefix is called the noun stem.

The stem remains constant, while the prefix may change, for example with the stem -*khulu* the following nouns may be derived:

Table 3.3 Different nouns derived from a stem

ikhulu (a hundred)	*amakhulu* (hundreds)
isikhulu (a person of note)	*izikhulu* (persons of note)
ubukhulu (size)	

From the above it is clear that, although the stem remains the same, the different class prefixes give different semantic contexts to each noun. Furthermore, as Bosch and Taljaard (1993, 1988) notes, these different class prefixes place each one of these nouns into a different noun class, even if the stem remains the same. Today, it is almost impossible for a non-mother tongue speaker to know to which class a certain noun belongs, because the majority of classes contain such a diversity of different nouns. The different noun classes are indicated in Table 3.4.

Table 3.4 *The different noun classes*

Class	Prefix	Meaning	Example
Class 1	*umu-*	Persons	*umuntu* (person)
Class 1a	*u-*	Terms of relationship, proper names	*ubaba* (my father)
Class 2	*aba-*	Plural of class 1	*abantu* (people)
Class 2a	*o-*	Plural of class 1a	*obaba* (fathers)
Class 3	*umu-*	Natural phenomenon	*umuthi* (tree)
Class 4	*imi-*	Plural of class 3	*imithi* (trees)
Class 5	*ili-*	Miscellaneous	*ilitshe* (a stone)
Class 6	*ama-*	Plural of class 5, collectives	*amatshe* (stones)
Class 7	*isi-*	Implements, miscellaneous	*isitsha* (a plate)
Class 8	*izi-*	Plural class 7	*izitsha* (plates)
Class 9	*in-*	Animals, abstracts, miscellaneous	*inja* (dog)
Class 10	*izin-*	Plural class 9	*izinja* (dogs)
Class 11	*ulu- or u-*	Long objects, miscellaneous	*uthi* (stick)
Class 14	*ubu-*	Abstracts, collectives, NO plural	*ubukhulu* (largeness)
Class 15	*uku-*	Infinitives from verb stems	*ukubana* (to see)
Class 16	*pha-*	Used to form locatives (no longer active noun class)	*phandle* (outside)

3.3.2 The system of concords

The class prefix is not only important because it indicates the classes to which the different nouns belong, but also because it links the noun to other words in a sentence (Bosch and Taljaard, 1988, 1993). This is done by means of a concord derived from the class prefix of the noun. This concord is then prefixed to the verb in the sentence.

The system of concordial agreement is important, because it forms the basis of the whole sentence structure of Zulu. For example in the sentence: "The woman washes the plate", the word for "the woman" in Zulu is *inkosikazi*, for 'wash' it is *–geza*, and for 'plate' it is *isitsha*. The subject noun *inkosikazi* must now be brought into concordial agreement with the verb root *–geza* by means of a subject concord *i–* (derived from class prefix *in–*): *Inkosikazi igeza isitsha.*

3.3.3 Other grammatical structures

Not only are concords derived from class prefixes of nouns, but **pronouns** and nominal stems as well. For example:

Isitsha sona lesi umbala waso muhle.
The colour of this plate is beautiful.

In the above example, *lesi* is a demonstrative that are found in Class 7 (the isi- class), whereas the pronouns *sona* and *waso*, are formed by means of the following, as indicated in Table 3.5:

Table 3.5 Formation of pronouns

Class	Word (s) referred to	Pronoun stem	Additional prefix/suffix	Pronoun formed
Class 7 (isi-)	isitsha	so-	absolute pronoun (-na)	sona
Class 1 (umu-)	umbala / muhle	-so	possessive concord (wa-)	waso

The **verb** in Zulu has one important characteristic, *viz* that it consists of a root or a radical that carries its basic meaning (Bosch and Taljaard, 1993). However, the root by itself cannot be used, so one needs to add certain prefixes and suffixes for the root to have a functional value.

The different suffixes (also called verbal extensions) are added to the root in order for the verb to be applied in different contexts. When a suffix is added to the root, it becomes a verb stem. For example, the root *FUND* has the basic meaning 'learn', and by suffixing different extensions, it acquires different meanings:

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

Verb stem and suffix (Zulu)	Meaning (English)
-FUNDa	learn or read
-FUNDela	learn for
-FUNDile	have learned
-FUNDisa	cause to learn/teach/instruct
-FUNDana	learn together
-FUNDisana	teach one another

As indicated in the above example, the verb is conjugated when represented by the different moods (indicative, imperative and subjunctive), tenses (present, future and past) and other verbal forms in which the verb stem may be used to appear as a complete word.

3.3.4 Term creation in Zulu

The creation of new terms is arguably the most efficient way whereby the scientific, educational and technical demands of a language such as Zulu can be addressed. But for terms to be created, the Zulu language has to draw upon internal and foreign sources. Among the most important methods of term creation used in Zulu that draw on internal sources aresemantic shift, derivation from verbs and idiophones and compounding. Terms drawn from foreign sources are generally known as loan words or adoptives (Wilkes and Nkosi, 1995). As with any other language pairs (for example English-Spanish or German-English), numerous problems are encountered when trying to retrieve information from different languages. The difficulty mostly lies with a 'rich' language like Zulu-that reflects its origin, history, mythology, exploits, legends, wisdom lore and world views-not all words are found in the dictionary, which makes it difficult to establish a complete word index to use. Some of the grammatical influences will now be explained below to clarify any difficulties that might exist.

3.3.4.1 *Term creation by means of semantic shift*

There are words that have originated because of a semantic shift. This means that the existing meaning of a word requires an expanded or modified meaning, to name a new (mostly related) concept. For instance:

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

Zulu word	English meaning	Modified English meaning
umnyango	door	department
isandla	hand	handwriting
bulala	kill	cancel
khala	cry	ring

One clearly sees that it would be difficult in the CLIR process to match query words to the headwords in the dictionary, if the link between the modified meaning and the original meaning is not made.

3.3.4.2 *Term creation by means of compounding*

Several new terms are created in compounding two or more words into a single word with a meaning that very often reflects the combined meanings of the constituted parts. For instance:

Table 3.8 *Creating new terms by combining different words*

Zulu word	English meaning	Meaning of the compound
umabonakude	television	*u-ma-bona* (see) + *kude* (far)
umakhalekhukhwini	cellular phone	*u-ma-khala* (cry/ring) + *ekhukkwini* (in the pocket)
impumalanga	east	*im-phuma* (come out) + *ilanga* (sun)
isidakamizwa	drugs	*isi-daka* (intoxicate) + *imizwa* (senses)

Compounding makes English-Zulu CLIR very difficult, because the compounded meaning is difficult to match to dictionary entries if the compounded meanings of the

different words are not linked to one another. It is not always the case that compound words are found in the dictionary itself.

3.3.4.3 Term creation by means of deideophonisation

Deideophonisation is a process that deals with the coining of terms from sounds that can be associated with the object or action that has to be named. This involves adding a prefix to the idiophone that depicts the action or the sound. For instance:

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

Idiophone (without prefix)	**Zulu word (with class prefix)**	**English meaning**
bhamu (sound heard when a gun is fired)	*isibhamu*	gun
thuthuthu (sound made by motorbike)	*isithuthuthu*	motorcycle
vungu (sound made by strong wind)	*isivunguvungu*	gale wind

Because most of these terms are deduced by means of cultural interpretation, it could affect retrieval results negatively. This is mainly due to the different interpretations one would have for the different sounds made by different elements.

3.3.4.4 Term creation by means of adoptives

Many new terms in Zulu exist because they were adopted from other languages (notably from English and Afrikaans). Loanwords may differ significantly in their written form from the original words, but they sound very much the same. For example: Zulu words and English words are similar (but not identical) to each other, like *imotho–motor* (Doke et al., 1990). The loanwords may not always be listed in translation dictionaries. Many of this kind of loanwords are adapted to the phonological system of Zulu (like its syllabic structure), before they become an integral part of the Zulu lexicon. They are also known as Zululizations.

The following are a few of the many adoptives found:

Table 3.10 *Loanwords in Zulu*

Zulu word	English meaning	Zulu word	English meaning
istradi	Street derived from Afrikaans 'straat'	*imenenja*	manager
imoto	Motor/car derived from Afrikaans or English 'motor'	*ihambega*	hamburger
itafula	Table derived from Afrikaans 'tafel'	*ikhompiyutha*	computer
ithikithi	Derived from English 'ticket'	*ithemperesha*	temperature

While most of the traditional loanwords were channelled to class five (ili–) in Zulu, many of them are now channelled to class nine (in–), despite the fact that they have a class prefix connotative of class five.

Many loanwords, however, (especially place names), have been adopted with only a class prefix added to the original form. For example *iBeitbridge* and *i-Amazon*. The use of the hyphen in I-Amazon is to separate the two vowels coming together with a glottal stop between them. (See Table 5.1, Rule 8).

3.3.4.5 Term creation by means of deverbalisation

Although a very productive strategy of term creation in the African languages entails the process of deverbalisation, whereby nouns are derived from verbs according to certain well-known derivative rules it may have a negative impact on retrieval results. The following are a few examples:

Table 3.11 *Creating new terms by deriving verbs from nouns*

Zulu verb	English meaning	Derived Zulu word	English meaning
-fundisa	teach	*umfundisi*	preacher
-zala	give birth	*inzalo*	(financial) interest
-thatela	gather what has been stored by people	*intatheli*	newspaper reporter

3.3.4.6 *Term creation by means of loan-translation*

Loan-translation means that Zulu words are a literal translation of the meaning of the same word in the donor language. For instance:

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

Zulu word	English meaning (literally)	English meaning
Indizamshini	fly machine	aeroplane
izithelo zethini	fruit of the can	canned fruit
isikhwama sokuyothenga	a bag of to go and buy	shopping bag

3.4 Dictionaries

When it comes to dictionaries, no other indigenous language in South Africa can really compare to Zulu. The number and variety of dictionaries in the Zulu language far exceed that of any other indigenous language in this country. Zulu is also the indigenous language with the oldest lexicographic history in South Africa, as it has the distinction of being the first indigenous language in which Jacob Döhne published a dictionary in 1857 (Wilkes and Nkosi, 1995). Most of the methods Döhne used (below) are still approved today:

- All the words are arranged under the root or stem letter;
- Derivatives are arranged according to their roots that show their etymology; and
- The dictionary is enriched with idiomatic sentences and phrases, illustrating the use of words.

A dictionary usually provides general translations, and it may have problems regarding aspects such as missing words, missing word forms and a lack of proper nouns (see Section 5.3 for the errors encountered with dictionaries). But the availability of bilingual dictionaries for many language pairs and the simplicity of using the dictionary for translation make it a good choice for Cross Language Information Retrieval.

3.5 Chapter Synopsis

This chapter was included in this study to sketch a background to an understanding of Zulu as an indigenous language, as well as to Zulu morphology. This is necessary to analyse the errors as experienced in the empirical study (described in Chapter 4) to improve retrieval performance.

This study is important to the transformation in South Africa, where language is a highly sensitive issue. It is hoped, by allowing people to use the computer in their mother tongue, it will stimulate pride in their language, as well as enabling them to learn in their mother tongue (made easier because of this). It is true that translating Zulu to English and vice versa does not remove all barriers to computer access, but it will help to eliminate some of them.

The next chapter will present a summary of the empirical work done on Zulu-English CLIR, describing the processes, the key qualitative results and findings of these results. The problems and possible solutions will be addressed in Chapter 5.

Chapter 4

To laugh often and much; to win the respect of intelligent people and the affection of children; to earn the appreciation of honest critics and endure the betrayal of false friends; to appreciate beauty; to find the best in others; to leave the world a bit better, whether by a healthy child, a garden patch or a redeemed social condition; to know even one life has breathed easier because you have lived. This is to have succeeded.

Ralph Waldo Emerson

4 Results of previous empirical studies

4.1 Introduction

In the following section the key qualitative results and findings of previous studies done by Cosijn et al. (2002a, 2002b, 2002c, 2002d) will be presented and compared to results reached in similar investigations reported in the literature.

In addition, the process of manually translating the queries, and the reasons for doing so will be discussed. For an English-speaking person to access a database in Zulu, the query will be constructed in English, while the English query will be translated into Zulu. The Zulu query will then be run against the Zulu database. The resulting documents will be in Zulu. Since there are no large Zulu databases available, the reverse process was tested, as described in Section 4.3 (namely to put Zulu queries through an English language database).

As there were no Zulu-English bilingual translation dictionaries available in electronic format, part of a printed dictionary had to be retyped manually into a word-processing program to do the empirical tests. Due to various constraints (financial) and restrictions (copyright issues), it was decided to only create a monolingual word list (Zulu entries only) in electronic format (Cosijn, 2002c, 2002d). The dictionary used was the 1990 edition of the Zulu dictionary by Doke et al.

The following issues will be dealt with in this chapter:

- An overview of CLEF will be provided–what it is about and why it was used in this study.
- Why were only 50 CLEF topics used?
- The best matching method will be determined.
- A brief overview of the InQuery retrieval system will be presented. This is necessary to describe the test runs and experiments done with the queries.
- The topic of n-gram matching that was mentioned in Section 2.3.1.1 will be discussed in more detail.

4.2 *A brief overview of CLEF*

According to CLEF's website "the Cross-Language Evaluation Forum (CLEF) supports global digital library applications by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes" (Cross Language Evaluation Forum, 2003).

Futhermore, CLEF's aim is to create a community of researchers and developers studying the same problems–by organising system evaluation campaigns–and then to facilitate future collaborative initiatives between groups with similar interests. CLEF also attempts to establish strong links, exchange ideas and share results, with similar cross-language evaluation initiatives in the US and Asia (who works on other sets of languages). The final goal of CLEF "is to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness in the global marketplace" (Cross Language Evaluation Forum, 2003).

A roadmap of CLEF's work (Cross Language Evaluation Forum, 2003) in the past three years can be summarized as follows:

Table 4.1 *A roadmap of CLEF*

	CLEF 2002	CLEF 2001	CLEF 2000
Purpose	Five evaluation tracks tested different aspects of mono- and cross-language information retrieval system performance.	Three main evaluation tracks tested multilingual, bilingual and monolingual (non-English) information retrieval systems. There was also a special sub-task for domain-specific cross-language evaluation, and an experimental track testing interactive cross-language systems.	Three main evaluation tracks, and a special sub-task for domain-specific cross-language evaluation.
Details	▪ Multilingual Information Retrieval ▪ Bilingual Information Retrieval ▪ Monolingual (non-English) Information Retrieval	▪ Multilingual Information Retrieval ▪ Bilingual Information Retrieval ▪ Monolingual (non-English) Information Retrieval ▪ Domain-Specific Mono- and	▪ Multilingual Information Retrieval ▪ Bilingual Information Retrieval ▪ Monolingual (non-English) Information Retrieval ▪ Special task GIRT

	▪ Mono- and Cross-Language Information Retrieval for Scientific Collections–Amaryllis and GIRT ▪ Interactive Cross-Language Information Retrieval	Cross-Language Information Retrieval ▪ Interactive Cross-Language Information Retrieval	
Resources used	The CLEF test collection for 2001 consisted of SGML formatted newspaper and news agency documents for Dutch, English, Finnish, French, German, Italian, Spanish, Swedish, Russian, Portuguese, Japanese, and Chinese from the same period.	The CLEF test collection for 2001 consisted of SGML formatted newspaper and news agency documents for English, French, German, Italian, Spanish and Dutch from the same period.	The CLEF test collection for 2000 consisted of SGML formatted newspaper documents for English, French, German and Italian from the same period.

4.3 Methods and data used in querying the database

A test set containing 50 CLEF 2001 topics (topics C041–C090) was used in the original study, where all 50 topics were officially translated and run against the CLEF database as described in Section 4.4. For the manual analysis of the errors in this thesis, 35 of these topics were chosen by the researcher and translated with the help of ten mother tongue speakers. The 35 topics were used to provide a representative sample of both narrow and broad topics. As query keys, the words of the title and description fields of the topics were used. It is important to note that in the CLEF tests, proper names and other words not contained in the translation dictionary were translated by an n-gram matching method. This method is described in detail in Section 4.4.

In the instance where a request in the English language (the source language) was read into a database in an indigenous target language (Zulu, in this instance), the process followed (in Cosijn, et al., 2002a) can be described as follows:

Firstly, the English query was translated into Zulu (one pair), and this Zulu query was then matched to the database index (also in Zulu). This should be applied to all the necessary pairs. Matches could be made through various techniques, but because no suitable Zulu morphological analysers were available, a dictionary translation method had to be used (see Section 2.3.1). N-grams was used to match the Zulu search strings

with the inverted index. Since there are many morphological analysers available for English, it was trivial to match the English words in natural language to the translation dictionary entries. For each English word, a number of Zulu translations were found, some of which were correct and some of which were incorrect in context of the query (owing to the ambiguity of natural language). All these words were then matched against the inverted index. Since the inverted index is not normalised, approximate string matching between the query words and indexed words was necessary. By using n-grams, this was managed quite well.

4.4 *The retrieval system and test queries*

The *InQuery retrieval system* (Allan et al., 2000; Callan et al., 1995) was used in this study. InQuery is a best-match retrieval system that also allows retrieval of strict Boolean result sets. All result sets, whether agreeing Boolean conditions or best match queries, are ranked. InQuery is based on Bayesian inference networks and it supports a wide range of operators, including strict Boolean (AND, OR, NOT) proximity operators as well as various best match operators (Allan et al., 2000).

The InQuery query language provides a set of operators to specify relations between query keys. For the *sum*-operator, the system computes the average of query key (or sub-query) weights. The *syn*-operator treats its operand query keys as instances of the same key; for the keys linked by the syn-operator an aggregate document frequency is computed (Sperer and Oard, 2000).

By matching the individual Zulu words in the topics against the words in the monolingual electronic Zulu word list, the following approximate string matching techniques were tested: (1) digrams, (2) trigrams, (3) classified s-grams, (4) edit distance, and (5) LCS.

The first three are *n-gram matching* techniques. In n-gram matching, query keys and index entries are decomposed into n-grams, i.e., into the sub-strings of length *n* (Pfeifer et al., 1996; Robertson and Willett, 1998; Salton, 1989; Zobel and Dart, 1995).

N-grams are formed of the adjacent characters of words. The *classified s-gram (skipgram) matching* technique is a new n-gram matching technique that is described in more detail in Pirkola et. al., (2002a). In this technique, digrams are combined both of adjacent and non-adjacent characters of words. Digrams are classified into categories based on the number of skipped characters. Only digrams

belonging to the same category are compared with one another. In Section 2.2.1.1, the problem of words not always being translated from the source language into target language was discussed. One of the most widely known methods for handling these translation errors is to pass the untranslated words to a CLIR query (the final target language query) as such. In the instance of spelling variants however, a source language form does not match the variant form in a database index, causing loss of retrieval effectiveness. Therefore, alternative methods for translation needs to be applied–this is where n-gram matching (and other approximate string matching techniques), or transliteration based on phonetic similarities between languages–comes in as a useful method to find target language spelling variants for source language words (Pirkola et al., 2001).

When using n-grams in information retrieval, the search keyword in the query is used to match database index entries. The best-matching entries can then be added to the query words. These will be treated as a synonym list. By applying various statistical methods, the number of entries to be added to this synonym list can be limited.

Zobel and Dart's findings (1995) suggest that updating n-gram indexes is straightforward and the index size and retrieval time is acceptable. Their results also show that n-gram matching is more effective than matching based on phonetic coding. But, Pirkola et al. (2001) argues that n-gram matching from a CLIR perspective has a lower effectiveness than from an approximate string matching perspective. This is similar to the problem of translation ambiguity in CLIR (see Section 2.3.1.6). It is likely that a query structuring method–based on defining alternative translations as synonyms–may be effective in the instance of n-gram based name searching in CLIR. N-gram matching, together with query structuring could then be used in both monolingual and cross-lingual name searching.

For all n-gram matching techniques used in this study, the degree of similarity between topic words and the word list entries was computed using a string similarity scheme similar to the one in Pfeifer et al.'s research (1996).

Consider the following example (in CLEF Topic C067, see Appendix A) where the target database index contains the Zulu word *abantu* (people) and its stem *–ntu*, and the English query contains the word 'person'. A translation dictionary from English to Zulu might give the translation *umuntu* for person, which matches neither of the index entries. Yet, by using digrams the words can be matched.

The digrams for the three words are as follows:

abantu: {ab, ba, an, nt, tu}
ntu: {nt, tu}
umuntu: {um, mu, un, nt, tu}

The similarity between two n-gram sets A and B can be computed as the figure $| A \cap B | / | A \cup B |$ which, in effect, counts the number of joint n-grams and divides this by the number of all distinct n-grams in the two sets (Pirkola et. al., 2001). In the example above, for the match between *umuntu* and the stem *ntu* the result is:

$| \{um, mu, un, nt, tu\} \cap \{nt, tu\} | / | \{um, mu, un, nt, tu\} \cup \{nt, tu\} | =$
$| \{nt, tu\} | / | \{um, mu, un, nt, tu\} | = 2/5 = 0{,}4$

For the match between *umuntu* and *abantu* the following is obtained:

$| \{um, mu, un, nt, tu\} \cap \{ab, ba, an, nt, tu\}| / | \{um, mu, un, nt, tu\} \cup \{ab, ba, an, nt, tu\}| =$
$| \{nt, tu\} | / | \{um, mu, un, nt, tu, ab, ba, an\} | = 2/8 = 0{,}25$

Apart from the three n-gram matching techniques described above, which is based on non-metric similarity measures, other string matching techniques based on metric similarities involve *edit distance* and *LCS*. Edit distance is defined as "the minimum cost required converting one string into another" (Pirkola et al., 2001) Conversion includes character insertions, deletions and substitutions. For these two words, their LCS is the longest character sequence of the sequences that occur in both words.

In establishing which of the five procedures described above would produce the best results, five of the CLEF 2001 topics (see Appendix A) were used in a trial run. For topics C041 to C045 (Appendix A), there were 75 Zulu source words. Ten of these were proper nouns which were not matched, and therefore not considered in the calculations. The results were thus based on the remaining 65 source words. For each of these source words, the six approximate best matches were listed for each of the five procedures. It was then manually established which one of these six words was the correct match for the source word. If the first word were the correct match, a value of 1 was allocated, a value of 2 if the second word were the correct match, and so forth. If there was no match, a value of 7 was given. The results are provided in Table 4.2.

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

Digram							
1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
36	11	2	3	2	0	11	**65**
Cumulative	47	49	52	54	54	65	
LCS							
1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
38	6	4	3	0	0	14	**65**
Cumulative	44	44	51	51	51	65	
Edit							
1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
33	4	2	5	4	1	16	**65**
Cumulative	37	39	44	48	49	65	
Trigram							
1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
37	7	4	1	3	0	13	**65**
Cumulative	44	48	49	52	52	65	
Skipgram							
1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
38	10	3	2	1	1	10	**65**
Cumulative	48	51	53	54	55	65	

To show the performance level of the test queries *the original English queries were run as baseline queries.* They contained (as query keys) the title and description words of the CLEF topics (the test queries were formed on the basis of the same words).

The original English queries and *undisambiguated CLIR (test) queries* were flat sum-queries, where query keys (a, b, c, ...) were combined with the sum-operator:

#sum(a b c ...)

This was done, because in many CLIR studies the *Pirkola method,* (i.e. treating translation equivalents as synonyms and combining them by the InQuery syn-operator), has been demonstrated to perform well (Ballesteros and Croft, 1998a; Gollins, 2000; Meng et al., 2000; Pirkola, 1998; Pirkola et al., 2002b; Sperer and Oard, 2000).

Two types of syn-structured queries were formulated. In *syn1 CLIR queries,* all the translations of the three best matches of a Zulu topic word were combined with the syn-operator. For example, the three best matches and their translations for the Zulu word *ephathelene* are as follows:

phathelela: grip, tightly, hold, lay, hands, make, constant, reference
phathela: handle, carry, treat, mention
phathelana: concerned, connected, relate

The syn-statement is as follows:

#syn(grip, tightly, hold, lay, hands, make, constant, reference, handle, carry, treat, mention, concerned, connected, relate).

In *syn2 CLIR queries,* the translations of each three matches of the Zulu topic word were combined with the syn-operator. For example, for the word *ephathelene,* the syn-statements are the following:

#syn(grip, tightly, hold, lay, hands, make, constant, reference) #syn(handle, carry, treat, mention) #syn(concerned, connected, relate).

In both instances (syn1 and syn2), the syn-statements were combined with the sum-operator.

Cosijn et al. (2002b) evaluated the effectiveness of the test queries as precision at a recall of 10%, and an interpolated average precision of recall levels between 10%–100%. The Wilcoxon signed ranks test, in conjunction with a statistical program based on Conover's research (1980), was used to test the statistical significance of the difference between the performances of the test queries against the baseline queries. The Wilcoxon test uses both the direction and the relative magnitude of the difference of comparable samples–in this instance the baseline and test queries.

As saturation normally occurs around the first three translated matches, the correct Zulu word should be identified within a set of three words about 80% of the time. Each word in the Zulu translations of the query topics was matched to the base-forms in the electronic monolingual Zulu word list. Based on the five topics analysed (as described above) it was decided that the three best matches of the skipgram

technique should be used, which identifies the correct base form in 78% of all cases. For each source word, the three best matches were translated and these were to be treated as synonyms.

4.5 Zulu to English Translation

As there was no bilingual translation dictionary, the base-forms were identified from the monolingual word list, and were then manually translated back to English. Strict rules were followed: all senses were recorded, exactly as they would appear in the dictionary. Words used in an idiomatic expression were not included. Hyphens were replaced with spaces and apostrophes were deleted. Stop words were removed manually according to a list. The stop word list consisted of source and target language files, containing frequently appearing words (sometimes even non-informative function words) in Zulu, and typical stop words such as prepositions and pronouns (see Appendix C).

The following example illustrates this process. Firstly it is given as a dictionary entry (as it is printed), and then the resultant manual translation is provided:

-khala (isikhala, izihkala) n.
Opening (permitting of a through passage or vision, as an opening between the hills, trees, clouds); gap (through a fence or wall). [cf. *intuба.*]
Opportunity. [cf. *i(li)thuба.*] *ithuба lokusinda* (an opportunity for escape).
Temporal region (above the ridge of the cheekbone and below the temples). [cf. *inhlafuno.*]
Open space, glade. [cf. *i(li) бathu.*]

The entry above becomes:
-khala isikhala izihkala: opening, permitting, passage, vision, opening, hills, trees, clouds, gap, fence, wall, opportunity, escape, temporal, region, ridge, cheekbone, temples, open, space, glade.

For all CLIR queries (undisambiguated, syn1, and syn2) proper names were managed differently. Borrowed proper names in Zulu are usually prefixed, with the stems being in their source language forms, e.g., *eSiberia*. The capital letters are kept in the middle of the proper names. This simplifies proper name identification and handling in CLIR. The prefixes were removed from the Zulu proper names (single and

hyphenated words having a capital letter in the middle of words), and the unprefixed forms (e.g., *Siberia*), were used in the final queries. In Zulu, borrowed phrases (common nouns) are usually indicated by applying inverted commas, e.g., "computer virus". As such, both unprefixed proper names and borrowed phrases were passed to the final queries. The number of queries containing these two expression types was 34 and 6 respectively (some queries contained both). The number of queries that did not contain either one of these was 13 (see Table 4.3). Because proper names in Zulu are identical to their English forms, removing the Zulu prefix from the proper name resulted in satisfactory retrieval performance of the syn-queries. However, it is not always the case that proper names are identical in Zulu and English respectively. Proper name queries also contained many mistranslated keys, but owing to the syn-structure it performed quite well.

The results are presented in Table 4.3. As shown, the average precision of baseline queries is 34,3%, while that of undisambiguated CLIR queries is only 4.0%. Syn1 queries perform substantially better than undisambiguated queries. Performance difference between syn2 and undisambiguated queries is also clear.

Table 4.4 presents the relative performance of CLIR queries (i.e. with respect to baseline queries). For syn1 queries, it is 58,6%–62,7%, for undisambiguated CLIR queries only 10,8%–11,7%.

As described above, the prefixes were removed from the Zulu proper names. The stems that were English words were passed unchanged to CLIR queries. It was obvious that unprefixed proper names contributed to the good retrieval performance of the syn1 queries. In addition, the phrases indicated by inverted commas probably had positive effects. To test the effects of these untranslated English words, both the test and baseline query sets were divided into two subsets: (1) proper name queries–queries which had either at least one proper name or an "inverted comma key" (the latter ones were infrequent), and (2) non-proper name queries–queries which only contained Zulu-English translations. The results of these runs are presented in Table 4.3. As can be seen, proper name CLIR queries perform quite well in relation to proper name baseline queries, while the performance of non-proper name CLIR queries is very poor ((that is, 3,5% (Pr. at 10% R) and 1,4% (Avg. Pr.) respectively)).

The performance of all CLIR queries were statistically different at the levels of $p = 0.01$ and 0.001 from that of the baseline queries for both evaluation measures (i.e. precision at 10% recall and average precision).

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

Query type	*Pr. at 10% R*	*Avg. Pr.*
N=50		
Eng - Original English	54.8	34.3
Undisambiguated	5.9	4.0
% Undisambiguated /Eng	10.8	11.7
Statistical sign. level	0.001	0.001
Syn1	32.1	21.5
% Syn1/Eng	58.6	62.7
Statistical sign. level	0.001	0.001
Syn2	17.2	11.7
% Syn2/Eng	31.4	34.1
Statistical sign. level	0.001	0.001

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

Query type	*Pr. at 10% R*	*Avg. Pr.*
Proper name queries, N=37		
Eng - Original English	59,7	39,7
Syn1	42,2	28,6
% Syn1/Eng	70,7	72,0
Statistical sign. level	0,01	0,01
Non- proper name queries, N=13		
Eng - Original English	41,3	19,2
Syn1	3,5	1,4
% Syn1/Eng	8,5	7,3
Statistical sign. level	0,01	0,01

Approximate string matching provided relatively good results, but several types of problems were experienced (e.g. inflected word forms and paraphrased translations). The correct base-forms were not always top-ranked, and this caused ambiguity in the translation process. It was found, that although paraphrasing may give some good keys, the syn-structure does not help in these instances. The latter happened because of the disambiguation effect of the syn-structure ((important keys occur as single keys (outside the syn-statements)). Also see Section 4.4 for more detail on the syn-

structure of the queries. This implies that more weight is assigned to single keywords than to mismatched keys. The mechanical matching of running text to dictionary entries through approximate string matching works relatively well, but at a conceptual level there are serious problems. This includes strings that are incorrectly recognized (Lopresti and Wilfong, 1999). For example, the bilabial implosive– previously written as ɓ–is represented by b, as in *ubaba, ubudoda.* It is a very common error for systems to identify the ɓ for a 6 in the written form. Unfortunately, the pronunciation of the words cannot be taken into consideration, as it sounds the same in almost all instances. One then needs to examine the grammatical analysis of the word to make a correct identification.

4.6 Chapter Synopsis

In this chapter, the key qualitative results and findings of previous studies done by Cosijn et al. (2002a, 2002b, 2002c, 2002d) have been highlighted and compared to results reached in similar investigations reported in the literature.

The process of manually translating the queries, and the reasons for doing so, has also been discussed. Also, three n-gram matching techniques (digram, trigram and classified s-gram) have been discussed in more detail.

Furthermore, a brief overview of CLEF was presented, substantiating why it was used in this study and the 50 topics. The reader was introduced to the InQuery retrieval system, as well as how the test runs and experiments were done with the CLEF queries.

This chapter also showed that dictionary translation (supplemented with fuzzy matching) seems to be the only viable option in a South African context (Cosijn et al., 2002d). However, because of the unavailability of large-scale Zulu databases, it is not possible to test the English-Zulu process.

Previous research (Cosijn et al., 2002a, 2002b, 2002c, 2002d) outlines a process based on monolingual approximate string matching in Zulu, to identify the inflected query word forms (as indicated in the dictionary) to the Zulu inflected forms. This problem may be solved as soon as a morphological analyser/parser becomes available. Promising research is currently done by Bosch (1999), whereby a morphological parser for Zulu is being developed, which may enable language-specific information retrieval in Zulu.

Results obtained in research done by Cosijn et al. (2002a, 2002b, 2002c, 2002d) suggested that while the translation of English words into Zulu base forms is generally manageable, it sometimes creates problems of conceptual incompatibility between the English and Zulu languages that may be difficult to solve.

Present research in CLIR concentrates on languages with comparable vocabularies of terms such as technical and scientific terminology. This chapter has shown that this research, together with research from Cosijn et al. (2002a, 2002b, 2002c, 2002d) indicates new problems that will be encountered if the language pairs used contain disparate vocabularies. This increases the complexity of CLIR, and in this instance, the problem has severely increased, because the languages dealt with lacks resources for word form analysis. These problems are further investigated in Chapter 5, but techniques have to be found to research in CLIR.

Chapter 5

Continuous effort, not strength or intelligence is the key to unlocking our potential.

Liane Cardes

5 Problem identification and analysis

5.1 Introduction

This chapter will focus on the possible reasons for the relatively poor retrieval performance of the empirical studies, as described in Chapter 4. In this chapter, it will be argued that two factors limit the performance of the n-gram approach where the Zulu search strings were matched to the inverted index.

The first factor is about translation ambiguity. Countless words do not have a unique translation, and sometimes the alternate translations have very different meanings. By applying every possible translation, the set of possible meanings can be greatly expanded, because some of the translations are likely to introduce additional homonyms or polysemous word senses into the second language. For instance, when untrained users enter such short queries (sometimes only one word) that it is not even humanly possible to determine the intended meaning (and the proper query translation) from the available context.

The second limiting factor regarding retrieval performance is the lack of essential dictionary terms for the correct interpretation of a query. This may occur either because the query is about a technical topic outside the scope of the dictionary, or because the user has entered some form of abbreviation or slang that is not included in the dictionary. As dictionaries (specifically designed for query translation) are developed, the effect of this limitation may be reduced. Nevertheless, it is unlikely to be eliminated, because language use is creative and new terms continuously enter the lexicon.

5.2 An analysis of errors

A detailed explanation of the process applied in this particular study was provided in Chapter 4 and only relevant sections will again be discussed for the purpose of contextualisation.

Five CLEF topics were used as a test bed, where the individual Zulu words (in the CLEF topics) were matched against the words in the monolingual electronic Zulu word list through:

- digrams,
- trigrams,
- classified s-grams,
- edit distance, and
- LCS.

All of these are approximate string matching techniques. For each of these source words, the six approximate best matches were listed for each of the five procedures. It was then manually established which one of these six words was the correct match for the source word (Section 4.4).

As saturation usually occurs around the first three (of the six) translated matches of the five different methods for the five queries (as above described), the correct Zulu word should be identified within a set of three words about 80% of the time (Section 4.4). Each word in the Zulu translations of the query topics were matched to the base-forms in the electronic monolingual Zulu word list. Based on the CLEF topics analysed, as above described, it was decided that the three best matches of the skipgram technique should be used (which identifies the correct base form in 78% of all cases).

Through a manual analysis of the matching of the queries to the original database entries, and a comparison of the official translations to the mother tongue translations two types of problems were identified (figure 5.1). They are **dictionary problems** and **translation problems**. In Figure 5.1 the two types of problems are compared in terms of the number of errors occurring with each. To categorise the errors causing the relatively poor retrieval performance, 35 queries were manually analysed. By literally counting the errors on a word-by-word basis, 169 occurrences of translation problems (in eleven sub–categories) were found, compared to 89 dictionary-related problems (in three sub-categories). This amounts to 258 instances (in 398 words matched to the index word list) of either dictionary- (34%) or translations problems (66%) in these 35 queries.

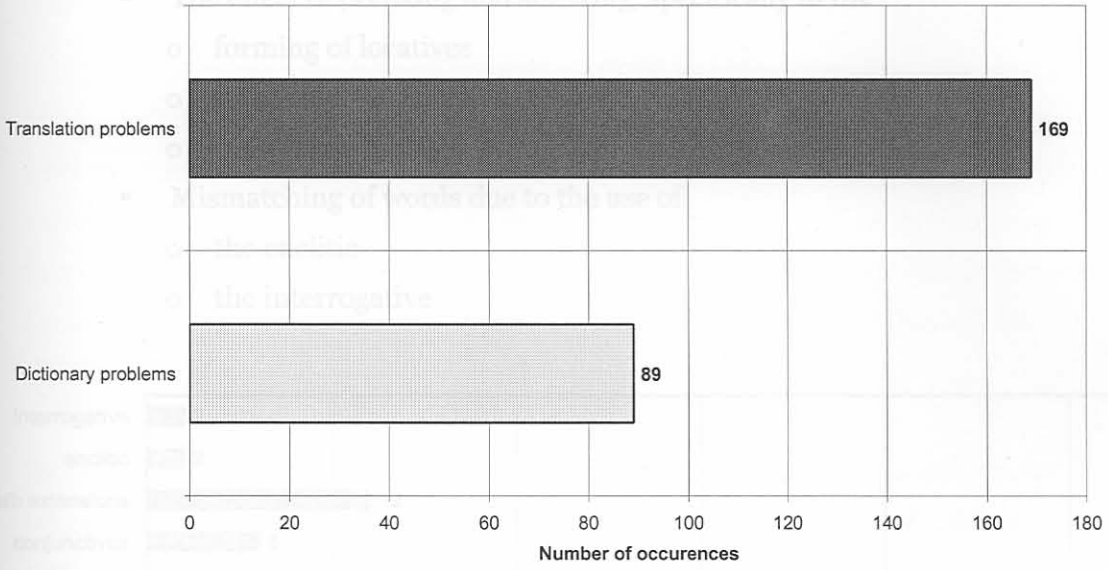

Figure 5.1 *A comparison between the dictionary problems and translation problems*

Typically, dictionaries provide several translations for a single source language word. Therefore the number of mistranslated keywords (i.e. the keywords that have incorrect meanings in the context of the topic in a CLIR query (the final translated query)) is usually high. Through the manual analysis of the queries, several **dictionary** problems were identified. These can be divided into:

- The orthographical rules of Zulu grammar;
- The frequent manner in which words are borrowed from English and Afrikaans;
- Zululised words; and
- The (mis)matching of proper names.

Apart from dictionary problems, numerous problems were also experienced during **translation**. These problems can be divided into the following:

- Paraphrasing, phrase translations and the effect of compounding;
- Changes to the noun and verb stems through word inflection, specifically
  - palatalisation
  - pre-nasalisation
  - the coalescence of vowels
  - vowel elision
- Homonyms and resulting mistranslations;

- The effect of prefixing and suffixing, specifically in the
  - forming of locatives
  - forming of conjunctives
  - verbal extensions
- Mismatching of words due to the use of
  - the enclitic
  - the interrogative

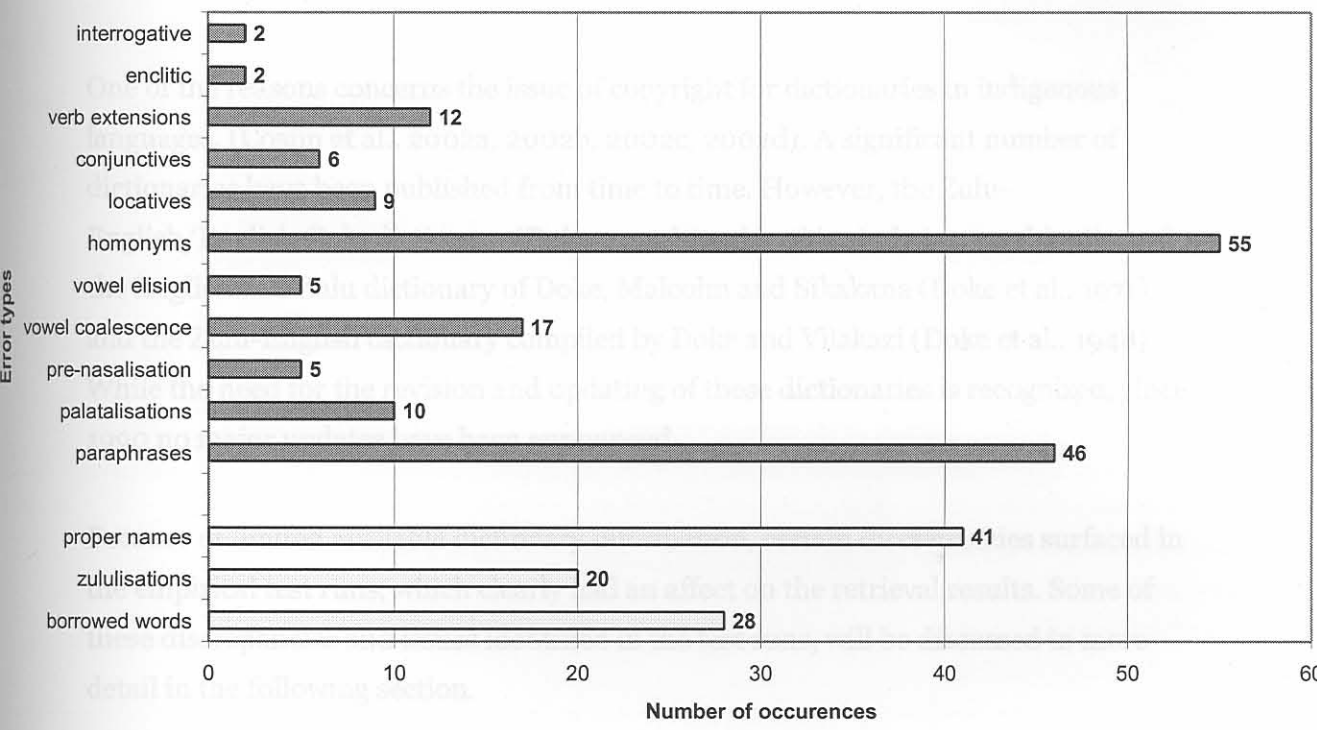

Figure 5.2 *A comparison of the types of errors encountered in the translation process*

Figure 5.2 compares the different types of errors encountered in the translation process with each other. Here, 398 words were taken from 35 queries and matched to the index word list. Of the 398 words, 258 words matched to entries found in the index word list, while there were no matches for 140 of the words. Of all the words (398), 262 were analysed and divided into two categories (dictionary and translation). Each had different sub-categories. The number of occurrences for each of the identified errors is indicated in Figure 5.2.

Each of the categories and sub-categories in Figure 5.2 will be discussed in more detail by using examples from the tables in Appendix A.

5.3 *Dictionary problems*

It is possible to improve the performance of CLIR by using more than one lexical resource, but during the research conducted for this study, other numerous problems were experienced with the available resources in electronic format. The number and variety of dictionaries in the Zulu language far exceed other indigenous languages in this country. However, currently there are no subject-specific or special dictionaries available in Zulu. This emphasised some of the errors experienced that will be discussed below.

One of the reasons concerns the issue of copyright for dictionaries in indigenous languages. (Cosijn et al., 2002a, 2002b, 2002c, 2002d). A significant number of dictionaries have been published from time to time. However, the Zulu-English/English-Zulu dictionary (Doke, 1990) used in this study is a combination of the English and Zulu dictionary of Doke, Malcolm and Sikakana (Doke et al., 1971), and the Zulu-English dictionary compiled by Doke and Vilakazi (Doke et al., 1948). While the need for the revision and updating of these dictionaries is recognized, since 1990 no major updates have been announced.

Because of limited available dictionary information, certain discrepancies surfaced in the empirical test runs, which clearly had an affect on the retrieval results. Some of these discrepancies and issues identified in the test runs, will be discussed in more detail in the following section.

5.3.1 Orthography

For a long time, Zulu was an oral language, and it is only recently (since 1848) that people have begun to document it (Wilkes and Nkosi, 1995). It is also only the third African language (besides Xhosa and Setswana) to appear in print. It is important to stress why a language needs to have a literary form. The lack of a written version can be costly to a language, as the integration with modern societies forces native people to adopt a recorded method of communication (Wilkes and Nkosi, 1995).

According to Wilkes (2001, 1988), the orthography of a language specifies a number of aspects about the writing of a language. Especially in how the speech sounds of a language should be written, as well as what the recognized words in the language are so that autonomous words are written separately from one another.

In 1934, the first Zulu orthography was officially recognized. Some of the most prominent changes occurred in 1950, when the Zulu orthography was revised and further amendments added in 1955. The latest edition, *Terminology and Orthography No. 4 for Zulu*, was issued in 1993. One of the implications of the updating and revision of the orthography is that certain discrepancies exist. Therefore, the following rules (Wilkes, 2001) pertaining to the writing of the sound system of Zulu are enforced, are indicated in Table 5.1:

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

Rule number	Rule description	Example in text	Comments
Rule 1	The bilabial implosive, previously written as *б* is represented by b.	*ubulala* (Table A1.1) *ubuzulu* (Table A3.2) *ubungeno* (Table A6.2) *ubukhanda* (Table A10.1) *ubuhlungu* (Table A13.1) *ubulima* (Table A18.2)	In the dictionary (Doke et.al, 1990) used for this study, some words were still written with a *б*. This made it difficult in some instances to correctly match the dictionary entries.
Rule 2	The devoiced bilabial explosive b will now be written as *bh*.	*imibhalo* (Tables A2.1, A2.2) *zebhola* (Tables A8.1, A8.2).	It does occur that mother tongue translators are not aware of changes to the writing system, and therefore still write words in the "old syntax".
Rule 3	The bilabial explosive in nasal compounds is written as *mb*.	*imbibika* (Table A34.2)	In most instances it is not easy to determine when the nasal sound influences the spelling of a word. This emphasizes the occurrences of palatalised words.
Rule 4	The voiced glottal fricative will be represented by *hh*.	*-hhalo* (Table A19.2) *-hholisaka* (Table 30.2) *-hhovini* (Table A35.2)	As with rule 2, the mother tongue translators are not aware of the changes to the orthography. However, this does not necessarily influence the resulting matches, but it should be taken into consideration.
Rule 5	The ejective alveolar affricate will be represented by *ts*.		This rule does not apply to the current study.

Rule 6	Where semivowels occur between vowels (meaning that two vowels come together without a glottal stop between them), a semivowel (*y* or *w*) may be inserted between these vowels.	*-nayiti* (Table A2.2) *-momoyi* (Table 3.2) *-shayisa* (Table A20.2)	This has a big impact on the identification of the stem used in matching the dictionary entry. It also frequently occurs in verb extensions. For more examples, see section 5.4.4.3.
Rule 7	The hyphen will be used as follows: • When a numerical is preceded by an inflected prefix. (Cardinal numbers are to be written without any concords preceding them. For example *umbuzo 10, isifundo 2*, but ordinal numbers preceded by *ngomhla ka* will be written with a hyphen. For example *ngomhla ka-10*). • To separate two vowels coming together with a glottal stop between them	*wezingcweti zebhola lomhlaba ngo–1994* (Tables A8.1, A8.2). More examples can be found in Tables A16.1, A16.2, A22.1 and A22.2. *i-UN* (Table A2.1) *i-El Nino* (Table A3.1) *i-Europe* (Table 11.1)	There is no specific rule about the use of the hyphen in a grammatical sense – it is mostly enforced for better readability in the text.
Rule 8	The apostrophe is used to indicate elision, but is not relevant to this study.		

The orthography of Zulu (including the above rules) is based on phonetic principles as with any other language. Issues regarding pronunciation, foreign acquisitions, the use of the hyphen, stress-indication and the use of capital letters will be discussed in the following sections.

5.3.2 Borrowed words

It is only recently that a language like Zulu has begun to be subjected to the westernised world. Because of this it does lack certain words that are standard in technologically advanced societies. For example, words associated with inventions like 'electricity', 'telephone', and 'computer'. Therefore, most of the borrowed words

found in the text are not found in the dictionary yet, although some of them have been included as new updates to dictionaries.

Zulu for example, has no known word for "computer virus", and because of this it has been borrowed from the English language. Even the sound of the letter 'r' has entered the Zulu language only through borrowed words such as 'radio', 'shirt' and 'computer'.

Zulu has two categories of borrowed words. Category one refers namely to words that are **adapted to Zulu pronunciation**, and category two refers to words that remain unchanged with only **class prefixes added to conform to the grammatical system**. In the first category, words are borrowed from another language, adapted to Zulu pronunciation according to the phonological system and eventually included in the dictionary. This is known as *Zululising* the foreign words (also see Section 2.3.1.4). An example of a borrowed word adapted from English, but which has not yet been included in the dictionary is *amakhemikheli* (Table A1.1). It sounds similar to the English equivalent 'chemical'. The Zulu word for chemical that do appear in the dictionary is *-phathalene* or *-thakiweyo*, and they are actually adjective forms.

Another example of a borrowed word is 'computer'. In Table A14.1 and Table A18.1 (Appendix A) the word is mismatched, because computer is Zululised to *ikhompiyutha*. Furthermore, the mother tongue translation does not have a word, so the rule for borrowed words suggests that a class prefix should be applied to the English word. This results in *'i-computer'*. Note that the class prefix is actually class five (*ili-*), but the *-i(li)* takes a silent form, as is the case with almost all borrowed words.

The same can be applied to 'virus'. The English equivalent is Zululised and a class prefix -ama is added to conform to the Zulu grammatical syntax. The result (*amavayirasi*) would not be found in the dictionary. Interestingly enough, the mother tongue translation do translate the word 'virus' into the following phrase: 'poison of the computer'.

In the following examples (Table 5.2), the Zulu words borrowed from the English language are displayed. Note, that although the spelling of Zulu words differs quite substantially from that of the English spelling, the pronunciation is very similar.

Table 5.2 Zululised borrowed words

CLEF topic	English word as in CLEF	Zulu equivalent
C042 (Table A2.1)	soldiers	*amasosha*
C051 (Table A8.1)	ball	*ibhola*
C060 (Table A15.1)	politics	*epolitiki*
C068 (Table A21.1)	synagogue	*isinagogo*
C074 (Table A25.1)	tunnel	*ithonela*
C075 (Table A26.1)	court	*ikantolo*
C075 (Table A26.1)	Germany	*iJalimane*
C078 (Table A29.1)	September	*uSepthemba*

It is interesting to note that the word 'synagogue' can also be paraphrased (see Section 5.4.1) into *nezindlu zesonto zamaJuda*, and when translated it becomes 'the dwelling place on Sundays of the Jews'.

Table 5.3 indicates the second category of borrowed words in Zulu. In this category, words are taken as is from the host language (in this instance English), and only a class prefix is added so that the word will conform to the Zulu grammatical system. However, in adding the class prefix some borrowed words are hyphenated as indicated in the table. There are no clear rules for the use of the hyphen. From Table 5.3, it can be seen that *iTour de France* does not take a hyphen, while *i-El Nino* does. This is because the hyphen in the instance of i-El Nino is used between the two vowels (see Table 5.1, Rule 8) to simplify the reading and writing of the word. The mother tongue translator even attempted to translate *Tour de France* to 'the journey that is endless'. Another example is the word 'computer'. In the instance of the official translation it is Zululised to become *ekhompiyutha*, while the mother tongue translation results in *i'computer'*. In this instance the borrowed word is put in inverted commas. This is done to simplify the reading and writing of the word, and to indicate that the word is indeed in its borrowed (English) form. There are no specific rules that justify the inconsistent use of the hyphen or inverted commas. It is the same with *igreen power* (compare with *i'green power'* or *i-green power* or *i-'green power'*) and *iMad Cow Disease* (compare *i-Mad Cow Disease* or *i-'Mad Cow*

Disease'). It may be assumed that the inverted commas are used to indicate that the class prefix is added to the phrase as a whole, and not just the first word. Grammatically, any of the written forms could be correct in terms of use.

Table 5.3 Borrowed words with a class prefix added

CLEF topic	English word as in CLEF	Zulu equivalent
C043 (Table A3.1)	El Nino	*i-El Nino*
C044 (Table A4.1)	Tour de France	*i-Tour de France*
C059 (Table A14.1)	Computer	*i-'computer'*
C086 (Table A33.1)	green power	*i-'green power'*
C088 (Table A34.1)	Mad Cow Disease	*i-'Mad Cow Disease'*

** In Zulu, borrowed phrases (common nouns) are usually indicated by using inverted commas. (See Section 4.5).

As indicated in Figure 5.2, there were 28 instances where borrowed words occurred, making it the fourth highest type of error to be encountered in the analysis of the words.

5.3.3 Proper names

Proper names in Zulu are usually found in class 1a, as indicated in section 3.3.1. However, it can be added that the singular and plural forms of the category of personal names in proper names in class 1a are indicated by an u- (singular) or o- (plural). Thus, uJames can become oJames (James and the others). This implies that all proper names will take the appropriate prefix, as indicated in Table 5.4.

Table 5.4 *Proper names in Zulu*

Category of proper name	Example of Zulu proper name
Personal names	*uJames, uSusan*
Names of the Deity	*uNkulunkulu* (God), *uJesu* (Jesus)
Books of the Bible	*uGenesise, uEksodisi, uEzra*

Days of the week	*iSonto* (Sunday), *uMsombuluko* (Monday), *uLwesibili* (Tuesday)
Titles of people	*uDokotela Nkomo, uProfesa Dube*
Geographical names (not place names)	*i-Amazon, i-Antatika*
Names of organizations, historical events, and government departments	*iNhlangano yeZizwe* (United Nations), *uMkhosi wePhasika* (Passover), *iMfuduko Enkulu* (the Great Trek), uKhisimusi (Christmas)
Place names	*KwaZulu, eGoli, eManzimtoti, eThekwini*

As indicated in Table 5.4, if the proper name starts with a vowel, a hyphen is normally added between the prefix and the proper name. This is also true for the borrowed words found in Appendix A. For instance *i-El Nino* (Table A3.1) and *u-Indurain* (Table A4.1). Furthermore, it must be added that in the experiments conducted, the proper name was not matched through approximate string matching. To illustrate this, take 'Haiti' as an example: Haiti is not matched in the dictionary, but because it is a proper name, it takes the *lase-* in front of the stem. The reason for this is that, as a rule, most place names take a *lase-* or *kwa-* as a prefix, indicating 'the place of' or 'the country of'. 'US' though, are matched to *Melika*, since the word is actually found in the dictionary. For 'UN', the system matches the word to a paraphrased form 'the place of nations that is united' or *weNhlangano yeZizwe*.

It is also interesting to note that the proper name 'France' has a dictionary entry (*eFulansi*), but the mother tongue translation applies the rule for proper names by just adding a class prefix to the English word (*eFrance*).

A "stripping technique" was used, whereby prefixes and affixes were automatically identified and removed, thereby restoring the proper name to its original form (Cosijn et al., 2002a, 2002b, 2002c, 2002d). It was then possible to match the proper name directly to the database through a table lookup procedure to see whether the root of the word was spelled correctly. These words were recognized automatically by the system through the predetermined grammar rules (as explained in Sections 3.2, 3.3 and 5.1.2).

Apart from the dictionary problems, other errors were also encountered in the translation process. This was mentioned in Section 5.2, and will be discussed now.

5.4 *Error analysis of the translation process*

While locating translations in dictionaries, the following problems were encountered: missing words, proper name translation, spelling conventions and technical term coverage.

It was quite evident in this study, as the dictionary did not contain some of the required words, and it did not include all the possible meanings of the words it referred to. However, most importantly, it was impossible to resolve translation ambiguities by just using a dictionary. The problem is that if two possible translations were given for the query term, one would not necessarily know which translation the system would choose (Cosijn et al., 2002a, 2002b, 2002c, 2002d). Also, by using only a simple dictionary look-up, a high degree of ambiguity is introduced thereby making it difficult to identify translation equivalents.

Some of the more specific errors that were identified and indicated in Figure 5.2, will be addressed now.

5.4.1 Paraphrasing

According to the Collins Dictionary of the English Language (Hanks, 1983), paraphrasing is "an expression of a statement or text in other words, especially in order to clarify". It is common to use paraphrasing in indigenous languages like Zulu, because some of the technical and scientific terminology do not exist.

In our experiments (Cosijn et al., 2002a, 2002b, 2002c, 2002d), a sworn translator translated the CLEF English topics into Zulu, and a Zulu mother tongue speaker then examined these translations. To demonstrate the affect that the fundamentally different vocabularies of English and Zulu could possibly have on dictionary translation, mother tongue speakers (who were not translators) again independently translated 35 of the CLEF topics into Zulu. The first set of translations from English to Zulu was called 'official translations', and the second set of translations was called 'mother tongue translations'. In both sets of translations, a high incidence (18%) of paraphrasing were found (see Table 5.5). This can specifically be ascribed to the fact that English and Zulu do not really have comparable vocabularies, especially for technical terms. In Section 2.3.1.4, a study by Ballesteros and Croft (1996) on English to Spanish text and Spanish to English text, revealed a considerable loss in precision when queries were translated word-by-word, due to the absence of comparable vocabularies for the two languages.

Examples of paraphrasing have been listed below in Table 5.5. Some of the paraphrasing occurs in the official translation, while others in the mother tongue translation. In some instances, the paraphrasing was different in both sets of translations.

Table 5.5 *Paraphrasing in Zulu*

CLEF Topic	English word as in CLEF database	Zulu paraphrase	Direct translation back into English
C041 (Table A1.1)	pesticides	*amakhemikheli abulala zonke izifo ezinengozi*	Chemicals that put to death all diseases that are dangerous
C041 (Table A1.1)	pesticides	*umuthi wokubulala izinambuzane*	medicine that kills insects
C043 (Table A3.1)	weather	*nesimo sezulu*	the shape of the sky
C043 (Table A3.1)	temperature	*kwizinga lokushisa*	the process of becoming hot
C043 (Table A3.1)	rainfall	*nokunetha kwezulu*	the coming of wetness to the sky
C044 (Table A4.1)	tour	*emncintiswaneni wohambo*	the competition of the journey that is endless
C044 (Table A4.1)	reactions	*ukuphatheka kwabantu*	to deal with the feelings of the people
C046 (Table A5.1)	embargo	*ukuvalelwa kokushintshisana ngempahla*	the closing of mutual exchange of goods
C058 (Table A13.1)	euthanasia	*ukufa ngaphandle kokuzwa ubuhlungu*	death that is right of the senses that are painful
C063 (Table A16.1)	earthquake	*ukuzamazama komhlaba*	the shaking of the earth
C073 (Table A24.1)	referendum	*isicelo sokuthola umqondo wabo bonke abantu*	the application to find out the opinion of everybody
C081 (Table A31.1)	airports	*ezikhumulweni zezindiza*	the place of airplanes
C088 (Table A34.1)	'Mad Cow Disease'	*isifo samatele*	the disease of the hoofs of animals
C090 (Table A35.1)	exporters	*abantu abathengisa kumazwe angaphandle*	people that trade from the country to the outside
C090 (Table A35.1)	frozen	*eqandisiwe*	that is very cold like ice

More examples of paraphrasing can be found in Tables A8.1, A12.1, A19.1, A21.1, A28.1 and A33.1 (Appendix A).

From the above table, it seems as if the dictionary translation attempts to describe each of the words in phrases of three to four words, like temperature (process of becoming hot), or air pressure (the influence of the wind) and even rainfall (the coming of wetness to the sky). When the mother tongue translators were asked why the dictionary and mother tongue translation differ so much in terms the paraphrased words, the reply was: "we do not really speak like that".

These forms of paraphrasing are the result of directly attempting to match the Zulu words to dictionary entries. These are just a few examples that demonstrate the instances where no matches were made (for example *nesimo* and *lokushisa*) and where mistranslated matches occurred. In the instance of *nesimo*, the noun stem is actually *-mo*, and therefore no match was made. This also accounts for *lokushisa*, which is found in the dictionary beneath *ukushisa*.

Although there is no real Zulu word for 'corruption', the mother tongue translator indicated that the word *ngenkohlakalo* is generally used for corruption in the sense of "events or happenings of a criminal nature".

It is clear that the dictionary translation and mother tongue translation are different from one another. In Table A33.1 (Appendix A) for instance, the mother tongue translation for 'renewable' resulted in the following paraphrased form: 'that can be made new' (*angenziwa kabusha*).

In addition, both forms of translations paraphrased 'describing' from the original query description to 'that show the meaning of' (the original translation is *echaza ngokusetshenziswa*) and 'that joins together work' (the mother tongue translation is *ochaza ukusetshenziswa*). Interestingly enough, both Zulu matches are the same, but the English matches slightly differ.

One of the cultural aspects that play a role when moving towards cultural CLIR is illustrated in the instance where the words "green power" are mistranslated in the mother tongue translation. The mother tongue translation has nothing to do with electrical power, but rather power in the sense of strength, and then it must be of a green colour as well.

In Table A34.1, the query concerns Mad Cow disease, and detecting documents related to this sickness. In the mother tongue translation, however, the term "mad cow" is translated to "the disease of the hoofs". This is mainly because Zulus are aware of the disease, and they know it comes from animals with hoofs (and not only cows)–therefore the culturally correct translation. The dictionary translation again only applies the rule for borrowed words and adds the class six prefix to obtain "i-Mad Cow disease".

These examples in itself provide a framework for in-depth research into cultural translation and the correct interpretation of texts, as a great portion of a message is lost through the self-interpretation of texts without putting it into a desired context.

Mechanically, n-grams match most of the individual words quite well to that of dictionary entries. On a conceptual leve,l however, the result is not always good. This can be ascribed to the system of grammatical rules (see Section 3.3) that are applied to the text. When these rules are enforced in the different categories, it generate all the possible dependencies that are allowed on a conceptual level. This was evident in the experiments, as several differences are found between English and Zulu (both on grammatical and conceptual level). Furthermore, in terms of paraphrasing it was found that the association between two classes did not always contribute to the user's (in this case the translator) understanding of the text. This is one of the main problems of paraphrasing–it is a *mechanical* process of exchanging words and phrases for synonyms. In effect, it discourages careful consideration of the meaning of the text itself. Consequently, the translator may interpret the text differently, and as a result make small changes in the written text. In doing this, the translator actually changes the meaning of the original text itself. As a result, the amount of paraphrased text becomes too general or unfocused, and which was not always effectively translated or matched on a conceptual level.

5.4.2 Word inflection

The degree of word inflection varies from language to language. For a very dissimilar language pair such as Zulu and English, the difference is obvious and is necessary to be addressed to ensure a good quality translation. Therefore, a statistical n-gram matching approach was well able to address this issue.

In a study by Pirkola et. al (2001), they indexed the unrecognized words (more than 119 000 forms) of their Finnish test database as digrams (see Section 4.2 for

definition) and found "that the frequency of such digrams that act as suffixes or are part of suffixes is often high in particular at suffix positions". The following illustrates this: the *in*-gram is a genitive suffix and common at all positions of strings, but it is far more common at the last position (where it occurred more than 65 000 times), i.e., genitive position than at other positions. Pirkola's findings suggest "that n-gram based stemming in which the n-grams with high frequency at suffix positions would be down weighted, would benefit retrieval" (Pirkola et al., 2001). This is also true for Zulu, where the different verb-extensions can be assigned certain weighted values to improve retrieval results.

If the source language words appear in inflected forms, they cannot be readily translated, because they do not match dictionary headwords (which are in base forms). Zulu is a highly inflectional language (30% of the errors were because of word inflection), and matching the words in the running text to the dictionary entries is often problematic. The reason for this is the changes to the syntactic function of the word, thus being inflected for the singular and plural, for future and past tenses, as well as other tenses.

5.4.2.1 Palatalisation

In Zulu, bilabial consonants may not be followed by -w-. In some instances this semi-vowel is dropped, but in most instances the bilabial consonant gives way to a corresponding prepalatal sound. The process whereby words are changed to conform to certain inflectional forms are known as palatalisation (Bosch and Taljaard, 1988).

When the noun ends in *-bu, -bo, -phu, -pho, -bhu, -bho, -mu, -mo, -mbu, -mbo, -mpu* and *-mpo*, the sounds change as follows:

Table 5.6 *The effects of palatalisation*

CLEF topic	Sound change	Zulu word	Grammatical analysis (with source word underlined)
Co65, Co76, Co86	**b>tsh**	*ukusetshenziswa*	*uku + sebenza + izwa*
Co74	**b>tsh**	*ukugcotshwa*	*uku + gcoba*
Co76	**b>tsh**	*catshangwa*	*cabanga*
Co48, Co57	**ph > sh**	*zokukhishwa*	*za + uku + khipha*
Co44	**m > ny**	*okhonyeliswa*	*o + khomela + iswa*
Co57	**m > ny**	*abafunywana*	*aba + fuma + na*

Co60, Co64	**m > ny**	*kungavunyelwe*	*ku + nga + vuma + elwa + e*
Co46	**mb > nj**	*ukuvinjelwa*	*uku + vimba + elwa*
Co79	**mb > nj**	*ngezinjongo*	*nga + izim + banga + ongo*

Other sound changes that occur (but not as frequently as those indicated in the table), are **mp > ntsh**, **th > tsh**, **n > ny**, **nt > ntsh**, and **nd > nj**.

Approximate string matching in such complex cases does not result in any meaningful match. It is suggested that the only solution for this problem would be an effective morphological parser (Cosijn et al., 2002d).

Palatalisation occurred in at least nine of the queries analysed in Appendix A (indicated in Figure 5.2), amounting to 4% of the errors encountered.

5.4.2.2 *Pre-nasalisation*

A homorganic nasal compound consists of the *-n-* of the *in-/izin-* class and a following consonant, for example *-nt-* in *into*. Some of the nasal compounds in Zulu are not that noticeable in the practical orthography of Zulu (Bosch and Taljaard, 1988). The following examples will suffice:

Table 5.7 *The effects of pre-nasalisation*

Nasal sound	Occurrence in zulu word	English meaning
ph > mp	impi	army / war
bh > mb	imbuzi	goat
f > mf	imfene	baboon
y > ny	inyama	meat
sh > ntsh	intshe	ostrich
kh > nk	inkabi	ox
g > ng	ingozi	danger
s > ns	inso	kidney
hl > nhl	inhlwathi	python

Other nasal compounds include **ph > mp**, **th > nt**, **d > nd**, **z > nz**, **hl > nhl** and **qh > nq.**

Occurrences of pre-nasalisation were found in five of the query words analysed in Appendix A. This is indicated in Figure 5.2, which amounts to 2% off all errors encountered.

5.4.2.3 Vowel coalescence

The Zulu basic vowel *a* (usually found in the instrumental *nga-* and the adverbial formative *na-*), when followed by the vowels *i* or *o*, may at times coalesce to form *nge-* or *ngo-*, as indicated in Table 5.8.

Table 5.8 *The three principal cases of vowel coalescence*

a + i = e
a + u = o
a + a = a

The inflected word forms resulting from vowel coalescence (7%) were found in more than seven of the queries in Appendix A, as indicated in Figure 5.2.

5.4.2.4 Vowel elision

Apart from the grammatical elision of initial vowels (when forming vocative interjections or after adverbial formations), vowel elision occurs when two word groups are combined to form a new word group. There are two kinds of elision:

- Elision of the initial vowel of the second word, for example *lo* + *umuntu* = *lomuntu* (this person) or *leso* + *isihlalo* = *lesosihlalo* (that chair).
- Elision of the final vowel of the first word, for example *inkosi* + *enkulu* = *inkosinkulu* (big chief) or *bonke* + *abantu* = *bonkabantu* (all people).

Other examples specific to the different noun classes are shown in Table 5.9.

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

Class	Vowel elided	Zulu example	English meaning
1st person singular	ngi + -eba > ngeba	Ngeba imali.	I steal money.
1st person plural	si + -osa > sosa	Sosa inyama.	We roast meat.
Class 2	ba + -ala > bala	Abafana bala ukudlala	The boys refuse to play.
Class 5	li + -eqa > leqa	Ibhubesi leqa isango.	The lion jumps over the gate.
Class 7	si + -opha > sopha	Isilonda sopha kabi.	The wound is bleeding badly.
Class 8	zi + -ephula > zephula	Izigebengu zephula amfasitele.	The gangsters break the windows.
Class 10	zi + -onga > zonga	Izintombi zonga imali.	The girls save money.

Examples of vowel elision can be found in Tables A4.1, A12.2 and A13.2. The occurrences of vowel elision (2%) compared to other error types analysed are shown in Figure 5.2.

5.4.3 Homonyms

Homonyms are words that have the same form or sound, but the words have different meanings. In Zulu, the difference in meaning is sometimes only conveyed by tone (Canonici, 1995). Tone, described by Doke (1968) as "the sequence of musical pitch upon the syllables of words" is a very important element in Zulu. This is because tone in Zulu is semantic–it is a deciding factor in the meanings of words. It is difficult for a machine to recognize where tone is indicated in a word, since tone is usually created by means of an acute symbol.

The Zulu language has a relatively high occurrence of homonyms (more than 50 occurrences in over 90% of the queries analysed). Compared to the other problems experienced in the test runs, homonyms amount to 21% of all problems, which make this the most frequent error encountered in the analysis. Table 5.10 indicates some examples of homonyms found in the analysed queries, with up to three different meanings produced for a word, depending on where the tone is. This indication of tone would, however, not be detected by the machine during the translation process.

Table 5.10 Examples of homonyms

CLEF topic	Zulu word	Dictionary meaning (English)	Second meaning	Third meaning
C043	*-netha*	get wet, be soaked, leak	be sluggish, inactive, drowsy	
C043	*-zulu*	sky, lightning, weather	Zulu language and mannerisms, member of Zulu nation	intoxicating drink made from the juice from the heart of the ilala-palm
C043	*-hlaba*	earth, world, land	something of no value	
C043	*-phumela*	Cluytia Pulchella, a small scrub of the Euphorbia-family, used as a love-charm	after-effect	come out, rise for, leave for, discharge for **or** to go out and relieve nature.
C048	*-xolo*	bark of a tree	species of small veld plant	peace, quiet, calm, goodwill, forgiveness
C048	*-zwe*	nation, tribe, clan, state	rapidly spreading brain disease	
C051, C055, C063, C065, C084	*-bika*	report, announce	species of ant with white spots	
C057, C058, C060, C074, C086	*-zwa*	to perceive with the various senses and thus hear, listen, perceive, understand, taste, smell, sense, feel	sense of feeling, mark of pressure, deep hole, nerve	main rib or spinal of leaf
C070, C075	*-shona*	sink, go down, disappear, go out of sight	set of sun, go down of moon, and even die,	lose heavily, become poor, bankrupt, ruined

The noun *-su* in topics C055 (Table A10.1 and Table A10.2) and C056 (Table A11.1 and Table A11.2) has different meanings when the three different class prefixes are added. When the class prefix *ili-/ama-* is added, *-su* means "plan, method, scheme", or "ring on a cow's horn marking each calf's birth". However, with *isi-/izi-* added, *-su* becomes "stomach, belly, abdomen and appetite" and even 'pregnancy'. When *u(lu)-/izin-* is added to the word, it becomes "human stomach", "tough pliable object" or "soft goatskin cloak worn by women".

In experiments conducted in previous test runs, it was found that the syn-structure of the queries as above described, manages homonyms quite well (Cosijn et. al, 2002a, 2002b, 2002c, 2002d).

From the examples in Table 5.11, it can be seen how frequently mistranslated query keywords occur in s-gram matching.

Table 5.11 *Mistranslated words as found in several queries*

CLEF topic	Source word in text	Matched keywords	English meaning
C041 (Table A1.2)	*ekudleni*	*-nkulukundleni* *unkulukundleni* *onkulukundleni*	species of caterpillar, larva of psychid moth, species of encased grubs
		-nkulukundleni *unkulukundleni* *onkulukundleni*	species of caterpillar, larva of psychid moth, species of encased grubs
		ekuqaleni	in the beginning
C042 (Table A2.2)	*kwezwe*	*-zwezwe* *umzwezwe* *imizwezwe*	spur of a cock
		-zwezwe *umzwezwe* *imizwezwe*	spur of a cock
		-nkweza *inkweza* *izinkweza*	day
C048 (Table A7.2)	*zokuphuma*	*-phuphuma*	overflow, bubble over, be overcome with emotion, have a miscarriage
		-phuphuma *iphuphuma* *amaphuphuma*	species of tree used as a love charm
		-phuphuma *imphuphuma* *izimphuphuma*	overflow, superfluity
C079 (Table A30.2)	*ukuhlolisiswa*	*hlolisa*	cause to live at ease, attack by surprise, ambush

		-wulukuhle *isiwulukuhle* *iziwulukhu*	pouring out in bulk, mass expulsion, abnormally bulky object
		-holisaka *uhholisaka* *ohholisaka*	large sack, mealie bag

Table 5.11 also indicates that in Table A1.2 (Topic C041), the Zulu word *ekudleni* is matched with words that are nowhere similar to the same meaning. This is because *ekudleni* is formed from the word *-dla*, which means food, eating or feast. The word is formed as follows:

uku- + -dla+ -ini

The *uku-* becomes *eku-* because of the influence of *i-* from *ingane* (babies). Also, because of vowel coalescence *dla+ini* becomes *-dleni.*

In Table A2.2 (Topic C042), the word *kwezwe* is formed by *kwa + izwe*. This is not reflected in the dictionary translation process, which means that words in the running text are matched to the available keywords in the Zulu list in base form. The resulting matches differ from the original word *-zwe*, which actually means country/land, where *kwezwe* becomes "the country of".

The same occurs in Table A7.2 (Topic C048), where *zokuphuma* is matched to the shown dictionary entries. The keyword however, is formed through vowel coalescence (Section 5.4.2.3) from *za + uku + phuma*, and thus are not matched to the original word. The stem *-phuma* means "to exit, go out", and thus bears no relevance to the matched dictionary meanings indicated in the above table.

Another example is found in Table A30.2 (Topic C079) with the word *ukuhlolisiswa.* In this instance the match is made to *hlolisa* (attack, surprise), *-wulukuhlu* (abnormally bulky object) or *–hholisaka* (large sack). But, the real stem is derived from *-hlola*, which means "inspect, look into, explore, examine, test".

If the translation for inauguration is analysed (Table A25.1), an interesting and important observation is made:

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

Original translation	*ngokuvulwa*	opening
Mother tongue translation	*ngokugcotshwa*	anointment

In the instance of the mother tongue translation, it seems as if a completely incorrect translation were made, but when placed into context, it is clear that the translator did not take into consideration the context of the sentence, but rather the word itself. For the Zulus, an inauguration is usually associated with royalty. Only kings and queens are inaugurated and anointed as holy ones. For them it might be confusing to anoint a tunnel, but it is the translation for inauguration.

The word *ukugcotshwa* is also an example of word inflection (see Section 5.4.2), where the *-b-* became a *-tsh-*. In this instance, no match could be found, because the system will attempt to match *-gcotshwa*, when the noun stem is actually *-gcoba*.

More examples can be found in Tables A3.1 and A3.2 (CLEF topic C043); Tables A4.1 and A4.2 (CLEF topic C044); Tables A12.1 and A12.2 (CLEF topic C057); Tables A14.1 and A14.2 (CLEF topic C059); Tables A25.1 and A25.2 (CLEF topic C074); and Tables A29.1 and A29.2 (CLEF topic C078).

In Table 5.13, the resulting matches for the Zululised *amakhemikheli* (no such word exists in the dictionary) is shown. However, the Zulu word for chemical that do appear in the dictionary is *-phathalene* or *-thakiweyo*, which are actually adjectives of the word.

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

CLEF topic	**Source word in text**	**Matched keywords**	**English meaning**
C041	*amakhemikheli*	*kheli ikheli amakheli*	address of letter, mannerism, private character
		-kheji ikheji amakheji	cage, bird cage, transport cage
		-kheli umkheli abakheli	person who addresses a letter

To illustrate the process of matching homonyms to their base forms as found in the word index, the following example found in CLEF topic C076 (see Appendix A, Table A27.1 and Table A27.2) can be considered.

A searcher poses the following question:

"In what applications are solar power energy used or considered for future use?"

The typical search terms would be 'applications', 'solar', 'power', 'energy', 'considered', and 'future'. The remainder of the words will be discarded as stop words (see Appendix B for a complete stop word list). As no morphological parser was used in this study, let us assume that an English morphological analyser would normalise this to 'application', 'solar', 'power', 'energy', 'consider', and 'future'. The simple dictionary matches would provide the following information as listed below:

Table 5.14 *Proposed dictionary matches without a morphological analyser*

English	English base form	Zulu *
applications	application	*isithobo*
solar	solar	*-elanga*
power	power	*amandla*
energy	energy	*amandla/isidlakadla*
considered	consider	*cabanga*
future	future	*isikhathi esizayo*

* This is only a simplistic view, and does not take into account that more than one match would be made to the Zulu word.

For the next step in the process, the translated Zulu query has to be matched to the Zulu document. There are various options for matching, depending on the nature of the inverted index. This will be discussed as follows:

5.4.3.1 *Matching the exact forms of the individual words, as they appear in the running text.*

In the first instance, where the inverted index contains the exact forms of the words as they appear in the running text and the matching is simply done on the Zulu singular plus prefixes in the instance of nouns, inflected forms will be missed. For example, in the instance of 'application', the Zulu plural *izithobo* in the document will

not be matched to *izithobo* in the query. The same can be applied to verbs (e.g. -*cabanga*) used in particular contexts as indicated in Table 5.15.

Table 5.15 *Matching inflected verbs in different forms*

Inverted verb	English equivalent
kade ***ngikucabanga***	I have just been thinking about you.
kade ***ngicabanga*** *ngakho*	I have long been thinking it over.
kuphumelela ***abacabangayo*** *emfudweni*	Those who use their brains in study, get on.
wavuka wangasebenza ***ecabanga*** *ukuthi yiSonto*	When he woke he did not work thinking it was Sunday.

Every inflected verb shown in the above table will have problematical results in exact matching. This is because the inflected form is matched to a base form.

5.4.3.2 *Matching the limited normalised word form*

In the second instance, it might be possible to remove a number of prefixes or suffixes through a simple stemming procedure in a simplified morphological analyser. One particular problem that may occur with this particular approach is the difficulty in recognising prefixes, since prefixes may phonetically change because of adjacent letters. Another problem may occur with verbs. This problem can become much worse, due to the complexity of verbal inflections. For example, the middle syllable of the stem -*cabanga* is a bilabial *b* and in the passive mood (through palatalisation), it changes to *tsh* (Doke et al., 1990). The various tenses then become: -*catshangwa* (is considered), -*catschangile* (was considered) and *zokucatshangwa* (will be considered).

Another example is -*sebenza*, where the middle syllable of the stem is also a bilabial *b*. The various tenses in this instance then become: -*setshenziswa* (is used), *setshenziswile* (was used) and *zokusetshenziswa* (will be used).

5.4.3.3 *Matching the normalised inverted index to the dictionary entries*

The third approach is normalising the inverted index to the dictionary entries through n-gram matching between the text and the dictionary entries. This would imply that the search words in the source language are translated and normalised to the stem of nouns and verbs as is found in, for example, the Zulu dictionary of Doke

et al. (1990). Since the inverted index is similarly normalised, there should be a one-to-one match between search terms and inverted index items. The normalised forms of the Zulu translation in the example above then become:

Table 5.16 *The effect of normalisation*

Normalised English	Zulu translation	Normalised Zulu stem
application	*isithobo*	*-thobo*
solar	*-elanga*	*-elanga*
power	*amandla*	*-ndla*
energy	*amandla/isidlakadla*	*-ndla/-dlaka*
consider	*cabanga*	*cabanga*
future	*isikhathi esizayo*	*-zayo*

However, normalising by removing the prefixes and suffixes to reduce nouns and verbs to stems might result in ambiguity, as in the following instance:

Table 5.17 *Normalisation and ambiguity*

Zulu stem	Ambiguous Zulu matches	English meaning
cabanga	*cabanga* *ulu-cabanga*	think, reflect, suppose, consider, imagine cartilage at the end of the breast-bone

In the Zulu dictionary (Doke et al., 1990), singular forms, plural forms and stems for nouns are listed. In using this particular dictionary, ambiguity problems for nouns should be reduced. Verbs, however, are listed only as stems. In addition to possible semantic ambiguity in the stems of the search terms (as in the example of *-cabanga* above), it is also conceivable that n-gram matching between the running text and the dictionary might result in multiple possible matches between a specific word and different dictionary entries. It also yields a very low matching rate in the instance of complex inflected forms. One method for possible resolving this particular problem, would be to take the two highest-ranking items in the inverted index. Another approach would be to use multiple query words and structured queries (Cosijn et al., 2002a, 2002c, 2002d).

5.4.4 Prefixing and suffixing

5.4.4.1 Forming of locatives

The locative indicates place and can, according to the context in which the word appears, be translated as "at...", "in..." or "to...". There are three ways how the locatives of nouns are formed:

- By means of the prefix *ku-*:
  kubaba (to father), *kumfana* (at the boy)
  In these examples, the initial vowel is discarded and substituted by *ku-*.
- By means of *e-...-ini*:
  Umuthi > *emthini* (at/to the tree), *intaba* > *entabeni* (at/to the mountain).
  In these examples the initial vowel is replaced with *e-* and a suffixing *-ini*.
  When the noun ends in *-bu, -bo, -phu, -pho, -bhu, -bho, -mu, -mo, - mbu, -mbo, -mpu* and
  -mpo, the sounds change because of palatalisation as described in Section 5.4.2.1.
- By means of the prefix *e-* without the suffix:
  A number of nouns only take the *e-* in the locative without the suffix *-ini* being added. For example *eMpumalanga* (east), *ekhanda* (the head), *ekhaya* (a home), *esitolo* (a shop), *enkantolo* (magistrate's office), *ehlobo* (summer), *emini* (a day) or *eGoli* (Johannesburg).

Approximately 3% of the errors that occurred in the analysed query words are because of the forming of locatives (see Figure 5.2).

5.4.4.2 Forming of conjunctives

The adverbial formative *na-* may be translated to "with, together with" and is prefixed to the noun. Vowel coalescence takes place as described in Section 5.4.2.3. However, in certain instances *na-* (and, also) is used to join together two or more nouns in Zulu, and as such it coalesces with the succeeding word without altering the grammatical significance of the word. These two forms of *na-* should not be confused with each other.

2% of all the errors that occurred can be ascribed to the forming of conjunctives. In Figure 5.2 this is compared to the other analysed problems as encountered in the text.

5.4.4.3 *Verbal extensions*

Verbal extensions are one of the main characteristics of the verb in Zulu (Bosch and Taljaard, 1988). A verb stem may assume a whole series of different meanings just by suffixing a particular verbal extension.

The different verbal extensions are indicated in Table 5.18.

Table 5.18 *The different verbal extensions occurring in Zulu*

Type of verb extension	Zulu verb	English meaning	Zulu verb with extension	New English meaning
Applied (-el-)	-pheka	cook	-phekela	cook for
Reciprocal (-an-)	-thanda	love	-thandana	love each other
Causative (-is-)	-funda	learn	-fundisa	teach
Neuter (-ek-)	-dla	eat	-dleka	can be eaten
Passive (-w-)	-bona	see	-bonwa	be seen
Passive (-iw-)	-dla	eat	-dliwa	be eaten
Passive	-loba	write	-lotshwa[1]	be written
Passive	-sebenza	work	-setshenzwa[2]	do work
Intensive (-isis-)	-bambisa	hold	-bambisisa	hold very tight
Reduplicated applied (-elel-)[3]	-bona	see / look	-bonelela	look after / care for
Reduplicated verb stem	-hamba	walk	-hambahamba	walk a little
Reduplicated verb stem	-akha	build	-akhayakha[4]	build a little
Reduplicated verb stem	-dla	eat	-dlayidla[5]	eat a little

1 See Section 5.4.2.1 for rules of palatalisation.

2 In certain instances of palatalisation, the extension *-w-* does not immediately follow the consonants.

3 The meaning does not always reflect the correct meaning, but usually indicates an intensified action. Apart from *-elel-*, the extension *-ezel-* is also used to indicate a repeated action.

4 In the instance of disyllabic vowel stems an *-y-* is placed between the stems.

5 In the instance of monosyllabic stems an *-yi-* is placed between the stems.

At least 5% of the problems that were experienced were due to verb extensions (see Figure 5.2 for a breakdown of all types of errors encountered).

5.4.5 Mismatching

5.4.5.1 The enclitic

An enclitic is a formative that is neither a proper suffix, nor used independently. In Zulu there are three formatives considered to be enclitics: *-ke* (then), *-bo* (express insistence) and *nje* (just, merely, simply). These enclitics are suffixed to words (mostly verbs) and have a definite syntactical meaning.

5.4.5.2 The interrogative

Interrogatives may either be in adverbial forms or a suffix. The two forms of interrogative suffixes, as can be seen in the examples in Appendix A, is *-ni* (what?) and *-phi* (where?).

5.5 Conclusions and research findings

In taking this particular study into consideration, it must be emphasised that the translations were done only because no Zulu databases were available. It is indicated in Appendix A that the official translations differ significantly from the mother tongue translations, and it has considerable implications for CLIR and future research on CCIR. This particular study demonstrates how the translations (performed by the mother tongue translators) reflect the local culture of the translators–and how *they* interpret the text. The implications of this for CLIR on Zulu databases are evident. If it is assumed that the content of such databases are IK-related, it is suggested that the 'oral form of Zulu be used to capture the IK content, and not 'academic' Zulu. To overcome the cultural and linguistic barriers experienced in the translation process, the cooperation of native speakers are required to see to it that the original intended message is conveyed in a clear and accurate manner.

Furthermore, through the analysed query words and identified errors (Chapter 5), it was deduced that the Zulu translation aims to transfer every textual aspect. This includes the raw informational aspect, the emotional appeal (quite evident in Zulu) and definite cultural differences in terms of the English translation. This clearly emphasises the difference in communication patterns between Zulu and English.

The three different approaches towards matching the translated Zulu queries to the dictionary entries were outlined in Section 5.4.3.1 – Section 5.4.3.3. These approaches will have to be empirically tested on a corpus to establish which option provides better results. The possibility of applying a simplified morphological

analyser depends largely on the predictability of the use of prefixes and suffixes in the indigenous South African languages. This may differ for the various languages. Promising research are done in developing such a simplified parser for each of these languages, but for purely realistic reasons one may have to opt for non-normalisation in the inverted index of the target language documents (Cosijn et al., 2002a). This may result in low hits for the n-gram matching.

By combining content and metadata searching it might be able to reduce the number of irrelevant items retrieved (Cosijn et al., 2002a). The reason for this is twofold:

- First, by describing the content in detail and capturing additional properties of the content, it could result in a significant drop in the irrelevant items retrieved.
- Second, metadata can act as a filtering mechanism for the identification of the broad field of relevance for instance a search for a certain plant could be delimited to the field of agriculture, medicine or food.

5.6 *Chapter synopsis*

This chapter focused on the different problems experienced when conducting the empirical studies as described in Chapter 4. One of the main problems associated with dictionary-based CLIR is the abundance of mistranslated query keywords in CLIR queries (also known as *translation ambiguity*). This was evident in the conducted experiments discussed in Section 4.3 by Cosijn et al. (2002a, 2002b, 2002c, 2002d) on Zulu-English queries.

The problems experienced with the dictionary-based approach were discussed in detail. These problems were divided into two main categories–dictionary problems and translations problems, each with its own sub-categories of errors. Each of the mentioned categories and sub-categories of errors have been discussed in detail, as it formed an integral part to this study in terms of the detailed error analysis. Apart from the two categories (dictionary and translation problems), the errors can further be divided into two other categories–those concerned with problematic matching, and those concerned with culture-related issues. These two categories connect with the research questions answered in Chapter 6.

The chapter concluded with a summary on the most important research findings that will lay the foundation for future research in the CCIR field.

Chapter 6

Many of us grow up with the idea that mistakes are bad, linking our self-esteem with continued success. We become afraid of making mistakes. So in order to achieve success, we tend to avoid areas that may lie outside the apparent realm of our natural talent. In this perverse equation, the secret of success becomes avoiding failure, leaving much of our potential untapped. In order to reach our full potential to learn, we must accept and then transform anxiety and fear, relentlessly seeking accurate information on our performance. What used to be perceived as criticism now becomes a gift for constructive growth.

Michael Gelb, Tony Buzan

6 Towards Cross-Cultural Information Retrieval

6.1 *Introduction*

In acquiring an actual cross-cultural retrieval capability it will require more than learning new approaches of communicating information. It will require the learning of new approaches of indexing and describing information.

Transforming one's awareness (a sense of one's personal or collective identity that includes your attitudes, beliefs and sensitivity) makes it viable to naturally accommodate cross-cultural factors when encountering people of dissimilar cultures. Instead of applying rules of "culturally-appropriate behaviour" one should rather speak and behave naturally.

This chapter intends to identify those cultural factors by categorising the errors experienced in the analysis in chapter five, and highlighting the most important issues to be addressed in future research. If this is successfully dealt with, it is possible that cross-language information retrieval could give rise to cross-cultural information retrieval.

6.2 *Categorising the problems identified in the error analysis*

Through a detailed analysis of errors in Chapter 5, two main problems (each with its own sub-categories) were identified. They are dictionary problems and translation problems. A summary of these problems and their sub-categories are presented in Figure 6.1:

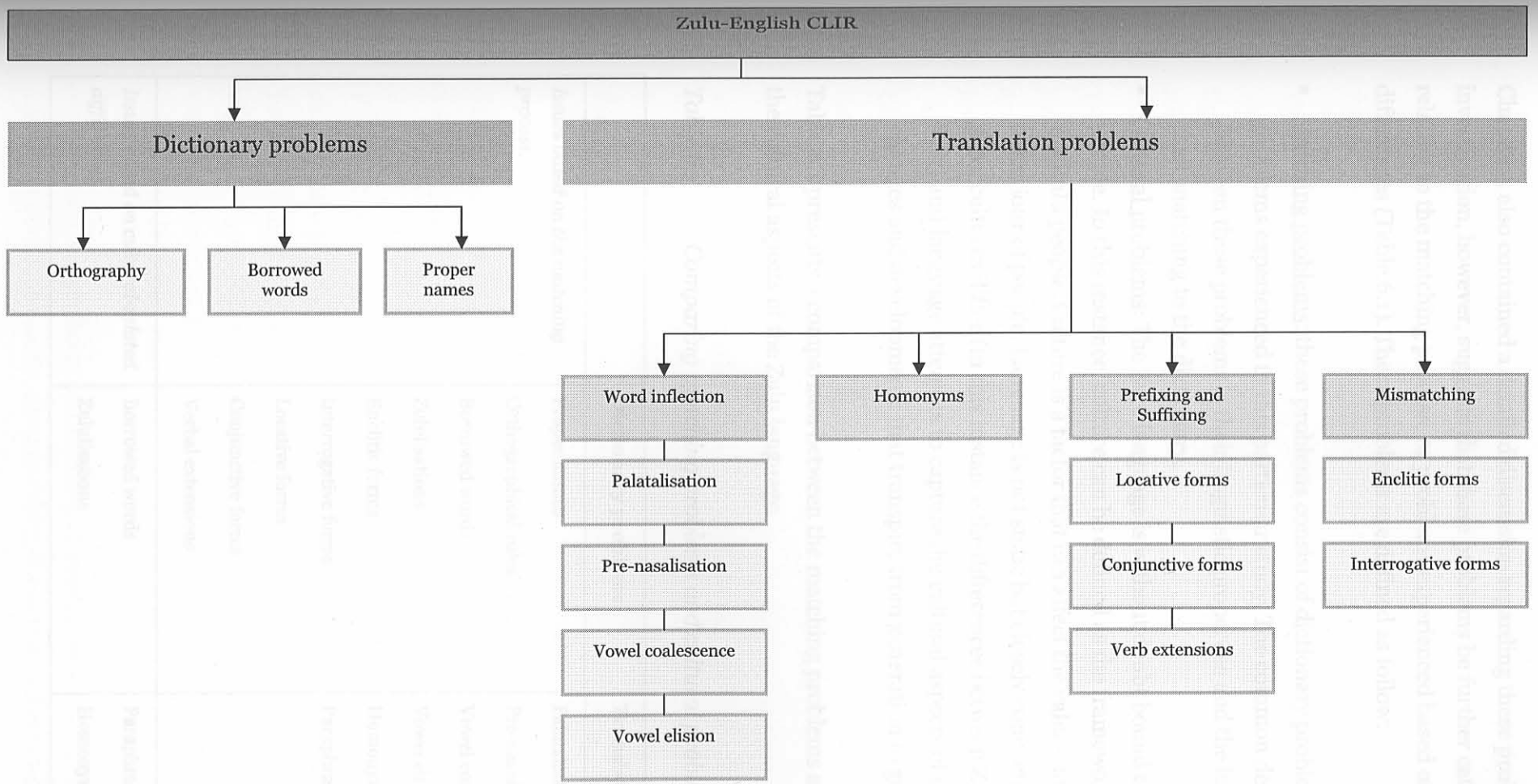

Figure 6.1 *A schematic representation of the analysis of errors*

Chapter 5 also contained a detailed discussion regarding these problems. Further investigation, however, suggests that these problems be further categorised as either relating to the matching process, or problems experienced based on cultural differences (Table 6.1). These problems are defined as follow:

- Matching problems: these problems consist of dictionary problems and translation problems experienced in this particular study. The common denominator between these problems is their linguistic properties and the impact it has on their matching to the dictionary.
- Cultural problems: The Zulu language is without doubt bound to the culture of its people. In this instance, culture can be defined as the framework for the lives of the Zulu people. Culture is a factor that can affect the values, attitudes and behaviour of people. Language is not static but closely connected with culture. Also, cultures differ (in this instance the differences between Zulu and English is vast) and language attempts to capture the cultural aspects of the society's changes and developments that transpire from generation to generation.

Table 6.1 presents a comparison between the matching problems and those related to the cultural aspects of the Zulu language.

Table 6.1 Comparing matching problems and cultural problems

	Dictionary problems	**Translation problems**
Issues based on the matching process.	Proper names Orthographical rules Borrowed words Zululisations Enclitic forms Interrogative forms Locative forms Conjunctive forms Verbal extensions	Palatalisations Pre-nasalisation Vowel coalescence Vowel elision Homonyms Paraphrasing
Issues based on cultural-related differences.	Borrowed words Zululisations	Paraphrasing Homonyms

From the matrix, one sees that some of the problems are not only matching related, but also cultural in nature. These issues will now be discussed.

6.2.1 Issues based on the matching process

The main cause for the following problems (both from a dictionary point of view as well as from a translation point of view) is that the words change so much in form that it becomes difficult and sometimes even impossible to match them to the original root words in the dictionary.

The dictionary problems stem from the orthographical rules applied to words that make it difficult for matching, while translation problems are mainly because of grammatical changes to words when certain occurrences are noted in specific instances.

On a matching level, it is evident that the Zulu language borrows extensively from English, Afrikaans and the Khoisan language. These borrowed words may take two different forms, namely words that are **adapted to Zulu pronunciation** and words that remain unchanged with only **class prefixes added to conform to the grammatical system** (Section 5.3.2). Even though the pronunciations are similar, the spelling of these Zululised words is vastly dissimilar from English words. This brought about that most of the borrowed words did not match any dictionary entries. Proper names are generally not translatable, which cause matching problems. This is because most proper names are not found in the dictionary. There are a few exceptions though, where words are found as Zululised forms (*iFulansi* for France in Table A15.1; *iYurobhu* for Europe in Table A21.1; and *eJalimane* for Germany in Table A26.1).

From a dictionary point of view two additional matching-related problems are the enclitic form (*-nje*) and the interrogative form (*-phi?* or *-ni?*). These words are normally added as suffixes to verbs or adverbs, but they are sometimes matched to dictionary entries. However, they do have a clear syntactical meaning. In the empirical experiments done, it was clear that these word forms were incorrectly matched to dictionary entries. This had a significant impact on the retrieval results.

Furthermore, the prefixing and suffixing of words (Section 5.4.4) also have a definite impact on matching results, since matching the words in their locative form, their

conjunctive form or as verbal extensions are problematic to achieve accurate results. This happens because these word forms are not words themselves, but they are mostly mismatched to parts of dictionary entries.

There are several orthographical rules that the Zulu grammar abides to, and which are indicated in Table 5.1. When some of these rules are enforced in the written form, certain problems arise. In some instances, it was difficult to look up the words in the dictionary, as they could not be found under the assumed entry. This was the case with a word like *-bhala*, which was actually found under the dictionary entry *-ɓala*. In pronouncing the word *-bhala*, one would never be able to distinguish the two different letters (b).

In the instance of paraphrasing, it seems as if the dictionary translation attempts to describe each of the words in phrases of three to four words. For instance temperature (the process of becoming hot); air pressure (the influence of the wind); and rainfall (the coming of wetness to the sky). These examples are indicated in Table A3.2. When the mother tongue translators were queried as to why the dictionary and mother tongue translation differ so much in the paraphrased word form, their reply was: "we do not really speak like that".

These forms of paraphrasing are the result of directly attempting to match Zulu words to dictionary entries. These are just a few examples that demonstrate where no matches were made (in the instance of *nesimo* and *lokushisa*) and where mistranslated matches were made. In the instance of *nesimo*, the noun stem is actually *-mo*, and therefore no match was made. This also accounts for *lokushisa*, which is found in the dictionary found beneath *ukushisa*.

Zulu is a highly inflectional language (30% of the errors were because of word inflection), and matching the words in the running text to the dictionary entries are often problematic. This occurs because of the changes made to the syntactic function of the word, thus being inflected for the singular and plural, for future and past tenses, and other tenses. The different palatilised and pre-nasalised forms are indicated and discussed in detail in Section 5.4.2.1 and Section 5.4.2.2. An applicable example is where the letter **b** in *-sebenza* changes to **tsh** to form *se**tsh**enziswa* (the causative *-is-* and passive *-w-* are added as verbal extensions).

Other word inflection forms described in Section 5.4.2, are vowel coalescence and vowel elision. Vowel coalescence occurs when the basic Zulu vowel *a* (usually found

in the instrumental *nga-* and the adverbial formative *na-*) coalesce to *nge-* or *ngo-* when followed by the vowels *i* or *o*. The latter occurs when two word groups are combined to form a new word. This occurs apart from the grammatical elision of initial vowels (when forming vocative interjections or after adverbial formations). Several other examples are presented in Section 5.4.2.3 and Section 5.4.2.4. The negative impact of these word inflections are evident in the results that were obtained, mainly because it was difficult to correctly identify the instances when words coalesce, or when palatalisation takes place. Because palatalisation occurs automatically, the matching process could not identify the source language words (appearing in their inflected forms). This is because the words that these inflected forms are being matched to, appear in their base forms as they appear in the dictionary.

6.2.2 Issues based on cultural differences

There is a vast difference between the Zulu and the English culture, which makes CLIR difficult. The Zulus have a different way of performing their daily tasks, and in expressing themselves. Through pictures, metaphors and idiomatic sentences they provide a more detailed explanation of their surroundings and feelings. This is not so for English, and because of these differences, some instances occurred where words did not match dictionary entries. Furthermore, the Zulu language frequently makes use of tone to indicate the meanings of words. Sometimes the exact same letter in words (spoken in a high or low tone) indicates different meanings. When the Zulu language is lacking a word for a certain concept in English, a new word is immediately created. This is evident in several of the Zululised words (Section 5.3.2). Not only is the paraphrasing of words difficult to match on a conceptual level, but the cultural aspects should also be taken into consideration in the matching process. This is mainly due to the richness of the Zulu language in describing words and actions. Paraphrasing is commonly applied to describe words or phrases in technical or scientific terms. Examples of paraphrasing are found in Table A1.1 (pesticides = *amahkemikheli abulala zonke izifo ezinengozi*) and Table A21.1 (synagogues = *kwezindlu zesonto zamaJuda*). More examples are discussed in Section 5.4.1. In the example presented, the word synagogue is paraphrased to "the dwelling place of Sundays of the Jews", but in Table A21.1 the mother tongue translation actually translates the word to *amasinagoge*, which is a borrowed form (Zululisation) of synagogue.

In Table 4.1 the word 'tour' occurs. The dictionary does not match the word, but leaves it as a proper name, and add a Zulu class prefix. The mother tongue translation though, translates the word to "the competition of the journey that is endless". Here the translator might have heard about the Tour de France, and placed it in his/her own context of an endless journey (possible because 3350km on a bicycle seems like a journey that would never end). If the word is left in its original form ('tour'), acceptable results can be obtained when the n-gram technique matches the word to its base form (and not the paraphrased form).

Although there is not really a Zulu word for corruption (Table A15.1), the mother tongue translator indicated that the word *ngenkohlakalo* is generally used for corruption in a sense of "events or happenings of a criminal nature".

Another example where cultural aspects play a role in moving towards cultural CLIR, is where the word "green power" (Table A33.1) is mistranslated in the mother tongue translation. The mother tongue translation has nothing to do with electrical power, but rather power in the sense of strength (and then it must also be green in colour).

In Table A34.1 the query concerns Mad Cow disease and retrieving documents related to this sickness. In the mother tongue translation, however, the term "mad cow" is translated to "the disease of the hoofs". This is primarily because Zulus are aware of the disease, and they know animals with hoofs are the cause of the sickness – therefore they provide the culturally correct translation. The dictionary translation again only applies the rule for borrowed words and adds the class 6 prefix to obtain "i-Mad Cow disease".

Through paraphrasing, it was found that the association between two classes did not always contribute to the user's (in this instance the translator) understanding of the text. This is one of the main problems of paraphrasing – it is a *mechanical* process of exchanging words and phrases for synonyms. In effect, it discourages careful consideration of the meaning of the text itself. Consequently, the translator may interpret the text differently and as a result make small changes in the written text. In doing this, the translator actually changes the meaning of the original text itself. This causes the amount of text paraphrased being too general or unfocused, which was also not always effectively translated or matched on a conceptual level.

In the instance of homonyms; the Zulu language is rich in the several different meanings it has for the same word. The deciding factor that determines the meaning,

is the tone applied in speaking the word, as well as the class prefix added to the word. It is difficult for a machine to recognize where tone is indicated in a word, since the tone is created by means of an acute symbol in most instances. For example: the word *-zwe* are found in several instances, and could translate to either 'nation' or "rapidly spreading brain disease". The same happens with *-ngena* in Table A6.2, where the word could either mean "the milking of a cow", "the purpose of taking a wife", "to enter the new year" or "to know intimately". Another interesting example is found in Table 10.2, where the word *-su* could translate to 'plan', 'method', 'scheme', or "rings on a cow's horn indicating the succession of birth" (This last translation is also a good example of paraphrasing).

In Figure 6.2, the issues (above discussed) relating to matching problems are compared to the culture-related issues. Interestingly enough, the 101 culture-related issues were experienced in only two instances (paraphrasing and homonyms), while the 109 matching related issues were experienced in nine instances (proper names, palatalisations, pre-nasalisation, vowel-coalescence, vowel elision, locatives, verb extensions, the enclitic and interrogative forms). The two instances (borrowed words and Zululisations) where the matching and cultural issues overlapped occurred 48 times.

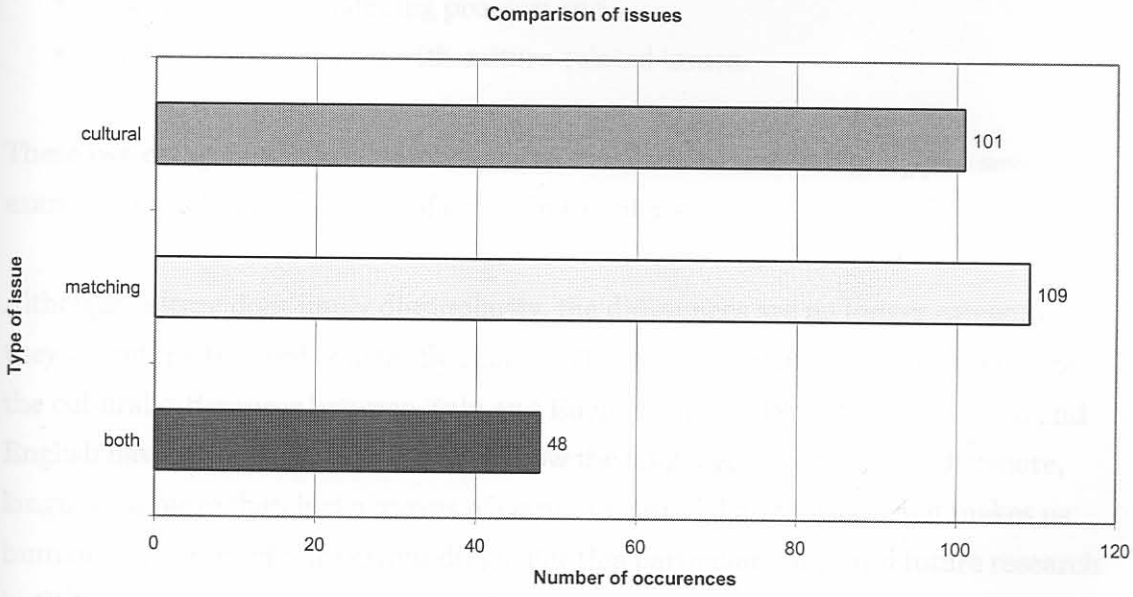

Figure 6.2 *A comparison of culture-related issues and matching issues.*

The reverse process that was tested in Cosijn et al's study (2002a, 2002b, 2002c, 2002d) may solve several of the matching problems experienced (as indicated in

Section 6.1.1 and Figure 6.2). However, the cultural issues need to be addressed in a different manner.

Most of the above mentioned problems encountered in the empirical experiments could be accounted to the morphological and syntactic structure of the Zulu language. It is clear that certain discrepancies exist in the comparison of mother-tongue translations and dictionary translations. The reason for this is mainly because of the contextual element associated with mother tongue speakers, and because dictionary translations are much more of a machine approach. These examples in itself provide a framework for in-depth research into cultural translation and the correct interpretation of texts, as a lot of meaning is lost through the self-interpretation of texts without placing it in a specific context.

6.3 *Chapter synopsis*

The concept of Cross-Cultural Information Retrieval was introduced in this chapter, and a short definition of what it encompasses was provided. Furthermore, this chapter also provided a concise summary of dictionary and translation problems (identified in Chapter 5) in terms of two new categories to which they could belong:

- problems in the matching process; and
- problems experienced with culture-related issues.

These two categories were identified, and a detailed discussion highlighted several examples found in the analysis of errors in Chapter 5.

Although culture does imply dissimilarity, the differences are no longer categorical; they are interactive and constantly change. This chapter attempted to acknowledge the cultural differences between Zulu and English. Dissimilar cultures like Zulu and English have various interpretations of how the language is applied. Furthermore, language is more than just a means of communication. Language is what makes us human, and aware of our surroundings. For this particular study and future research in CCIR, it is critical to be aware of Zulu behaviour, their beliefs and their values (stemming from personal and professional culture). Culture is the soul of the Zulu people. Through language, information can be passed on, one can learn from and connect with others, form and cultivate relationships, analyze, abstract and evaluate facts and concepts.

Language has several faces and layers, like the people who use it to communicate. An infinite number of possible grammatical sentences can be built with any linguistic inventory. Through the analysis of errors in Chapter 5 and Section 6.2 regarding culture-related issues, several examples indicated how the Zulu language is utilised to analyze thoughts, inform, dream, express emotion, convince, analise and confuse. The list of linguistic applications is endless. Also, the richness of the Zulu language proves how it evolved over time, and how it became part of the culture that defines it today. This is evident in the rhetorical speech, shifts in meaning, borrowed lexicon and structures, connotations, denotations, alliterations, metaphors, satire, parody, irony, registers, syntax and semantics.

The discussion regarding the cultural aspects will be used specifically to answer the research questions posed in Chapter 1.

Chapter 7

Effort only fully releases its reward after a person refuses to quit.

Napoleon Hill

7 Conclusion

The focus of this study was a detailed analysis of the errors encountered during Zulu-English CLIR, and why the results were not completely satisfactory.

In Section 1.2, the following main research question (MQ) was presented:

What were the main problems associated with the dictionary approach to Zulu-English CLIR?

In support of the main research problems, the following sub-questions (SQ) were asked:

SQ1: How successful were these particular experiments in Zulu-English CLIR?
SQ2: What reliable solutions can be implemented to address these specific problems?
SQ3: What can be done to contextualise language and culture in terms of CLIR?

The results of the empirical research (as discussed in Chapter 4) and conclusions reached in Section 5.5 will attempt to answer the questions as follows:

7.1 MQ: What were the main problems associated with the dictionary approach to Zulu-English CLIR?

The main problems identified in Chapter 5 (and which made Zulu-English CLIR difficult), were dictionary problems and translation problems. These problems were discussed in detail, and both were divided into sub-categories. With dictionary problems, there were several occurrences of borrowed words in the form of Zululisations and those that only took a class prefix (but kept in the original English form). Orthographical grammar rules in Zulu also made it difficult to retrieve some words in the dictionary.

The errors that occurred in the translation process were primarily paraphrasing, inflected word forms (palatalisation, pre-nasalisation, the elision and coalescence of vowels), homonyms, prefixing and suffixing (locative and conjunctive forms, verbal extensions), and mismatched word forms (the enclitic and interrogative forms).

These two primary problems can be categorised further into either problems related to matching issues, or problems related to culture-related issues. A detailed discussion was presented in Chapter 6, indicating the specific differences in each.

7.2 *SQ1: How successful were these particular experiments in Zulu-English CLIR?*

The results obtained during the empirical experiments done by Cosijn et al. (2002a, 2002b, 2002c, 2002d) were presented in Table 4.3. As indicated, the average precision of baseline queries is 34,3%, while that of undisambiguated CLIR queries were only 4,0%. Syn1 queries perform substantially better than undisambiguated queries. This must, however, be placed into the context of the relative performance of the baseline CLIR queries. Table 4.4 has indicated that for syn1 queries, the relative performance resulted in 58,6%–62,7%, and for undisambiguated CLIR queries in only 10,8%–11,7%.

As described, the prefixes were removed from the Zulu proper names and the English stems were passed unchanged to CLIR queries. Unprefixed proper names contributed to the acceptable retrieval performance of the syn1 queries. In addition, the phrases indicated by inverted commas probably had positive effects. To test the effects of these untranslated English words, both the test and baseline query sets were divided into two sub-sets: proper name queries and non-proper name queries. The results of these runs were indicated in Table 4.3. As can be seen, proper name CLIR queries perform quite well in relation to proper name baseline queries, while the performance of non-proper name CLIR queries is very poor, (that is, 3,5% (Pr. at 10% R) and 1,4% (Avg. Pr.) respectively). This means that queries that contained proper names or words in inverted commas were more accurately matched than queries that only contained the Zulu-English translation.

To categorise the errors causing the relatively bad retrieval performance, 35 queries were manually analysed. In literally counting the errors on a word-by-word basis, 169 occurrences (66%) of translation problems were found, compared to 89 dictionary-related problems (34%). This amounted to 258 errors found in these 35 queries. A summary of these errors is presented in Figure 5.2.

With approximate string matching, relatively acceptable results were obtained in matching the running text to the dictionary entries. However, on a conceptual level, Zulu-English CLIR was very poor due to the lack of technical terminology in Zulu.

Several of the search terms could not be translated directly to Zulu, as the language does not have single word translation equivalents for many of the technical and scientific terms. Although paraphrasing partly solved this problem, it still had a significant negative impact on retrieval results, which causes problems for Zulu-English CLIR. In addition, the poor retrieval performance can be ascribed to the system of grammatical rules (see Section 3.3) that is applied to the text. When these rules in the different categories are enforced, it generates all of the possible dependencies that are allowed on a conceptual level. This was evident in the experiments, as several differences are found between English and Zulu, both on grammatical and conceptual level.

7.3 *SQ2: What reliable solutions can be implemented to address these specific problems?*

It is proposed that the following solutions are investigated as possible solutions in improving results for Zulu-English. As no empirical data is currently available, it is not possible to say how much improvement could be made, but it is certainly worth investigating. The proposed solutions are:

- Query expansion;
- Applying metadata to describe the content;
- Applying normalisation; and
- Improving dictionary coverage to manage untranslatable terms

These proposed solutions will now be addressed.

7.3.1 Query expansion

Translation ambiguity is one of the primary hurdles that need to be resolved for Zulu-English CLIR results to improve effectively. Resources for CLIR may require a great deal of manual effort, as discussed in Chapter 4. Therefore, methods need to be obtained that capitalise on existing resources.

As indicated in the results in Chapter 5, there were three factors for the translation errors experienced:

- Unrelated terms added to the query terms;
- Failure to translate technical terminology (which is not often found in generalised dictionaries); and

- Failure to translate paraphrased terms, or poor translation thereof.

The first factor can be ascribed to the dictionary entries that (may) list several senses for a particular term, where each term has one or more possible translations. Apart from word-by-word translations having a tendency to be incorrect, poor translations also decrease retrieval effectiveness. Research by Ballesteros and Croft (1997) has indicated that query expansion can be successfully applied to significantly reduce translation ambiguity. Query expansion can be defined as "a set of techniques for modifying a query in order to satisfy an information need" (Selberg, 1997). Coverson (2000) describes query expansion as "the process of supplementing an original query with additional terms in order to refine a search and increase retrieval effectiveness". According to Selberg (1997), in most cases, terms are added to an existing query. However, query expansion also encompasses techniques for the reweighting of terms as well as the deletion of terms. By including more terms in the query, more context can be provided and may even further disambiguate translations. In particular, by including additional terms that have unambiguous translations themselves, a link can be established that may correctly indicate the context. Hull and Grefenstette (1996) indicated in their research that the retrieval performance achieved by manually translating the phrases in queries is not only significantly better than a word-by-word translation using a dictionary, but also more precise. Furthermore, Davis and Ogden (1997) indicated that in using a phrase translation dictionary, the performance of CLIR is also significantly improved.

However, the question remained: if the Zulu phrases were not found in the bilingual dictionary, how would the translator identify it in the query and translate it correctly? It would be unrealistic (for the translator) to expect a 'complete' phrase dictionary, or any 'complete' dictionary for that matter. New words and phrases are constantly created, especially in Zulu. Therefore, it is clear that the translator will always face the problem of phrase identification and translation thereof, no matter how complete the lexical resources. For instance, in the experiments it was indicated that 48 occurrences of paraphrased terms were found, which was almost 20% of the total number of errors analysed.

In the first analysis (described in Section 4.4) five queries were analysed (Cosijn et al., 2002b) and it was decided that the first three matched resulted in 80% correct matching. Future research might focus on the following questions: First, would it be possible to improve matching results if the query were expanded to six matches (and not only three as currently used)? Second, in the instance of expanding the query

matches, what would the (improved) effect be? In this specific experiment with Zulu-English CLIR, expanding the query words to take into account surrounding terms might improve results. However, based on the current (small) data samples, it is not possible to know for sure what the differences would be. Future research would have to investigate this in more detail with the aid of a bigger test sample.

In addition, future research should also investigate the viability of interactive query expansion (IQE) to improve retrieval results. Selberg (1997) refers to IQE as techniques (that encompasses relevance feedback) where the user has some interaction with the system in the query expansion process. The system usually suggests possible expansion terms and the user selects those they wish to add to the query. Studies (Coverson, 2000; Efthimiadis, 2000) have shown that interactive query expansion has the potential to improve retrieval effectiveness.

7.3.2 Applying metadata to describe the content

The Zulu language and the reasoning it reflects stem mainly from the cultural need for expression. This study concentrated on the cultural and linguistic problems that one faces when using the dictionary-based approach to CLIR. A practical solution that can be implemented to improve retrieval performance is the application of metadata (Cosijn et al, 2002a). In Milstead and Feldman's (1999) opinion, metadata is very important:

> While metadata has become a buzzword in the information business, the concept is important for both authors and seekers of electronic information. Used effectively, it makes information accessible by labelling its contents consistently. Metadata leaves a pathway for users to follow to find the information they need–all in one place. In invisible cyberspace, this is even more important than in a library where desperate users at least have shelves to browse.

The most current forms of multilingual access to information are inadequate to answer the needs of the increasing, diverse user groups from different cultural and linguistic backgrounds. The purpose of metadata is to describe the structure of the data for this existing information needs. Furthermore, any additional properties that the data might have can also be captured through categories specifically created for this purpose, to better understand the cultural concepts. Through these categories,

one can begin to address the cultural differences that exist in a language like Zulu. To make metadata accessible, however, it is suggested that it is in one language only. It is proposed that the metadata should be in English, for pure logical and realistic reasons (Cosijn et al, 2002a). If one has metadata in Northern Sotho, it would be useful for Northern Sotho information objects, but Northern Sotho metadata would not make Zulu documents accessible.

7.3.3 Applying normalisation

In Sections 5.4.3.1–5.4.3.3 different options of normalisation for matching words were described:

- The first option is where the inverted index contains the exact forms of the words as they appear in the running text. The matching is done simply on the Zulu singular plus prefixes in the instance of nouns. However, with this option inflected forms will be missed.
- The second option makes it possible to remove a number of prefixes or suffixes through a simple stemming procedure in a simplified morphological analyser. Unfortunately, it becomes problematic to recognise prefixes, since they may phonetically change because of adjacent letters. Another problem may occur with verbs, which can become much worse due to the complexity of verbal inflections.
- The third option is to normalise the inverted index to the dictionary entries through n-gram matching between the text and the dictionary entries. This implies that the search words in the source language are translated and normalised to the stem of nouns and verbs.

In most instances, it seems possible to remove a number of prefixes and suffixes through a simple stemming procedure in a simplified morphological analyser. This method has several benefits, since it reduces words to their base form. First, there is no need to be concerned about truncation or word inflection, since the different forms of the keywords are automatically conflated into the same form. In addition, retrieval performance will improve (especially recall), since a larger number of potentially relevant documents will be retrieved.

The above (normalising) options of matching the Zulu words to the dictionary entries have to be tested empirically on a corpus to establish which specific option provides

the better results. The possibility of applying a simplified morphological analyser largely depends on the predictability of the use of prefixes and suffixes in the indigenous South African languages (this may differ for the various languages). Currently, promising research is in progress regarding morphological parsers.

7.3.4 Improving dictionary coverage to manage untranslatable terms

The following table indicates some of the untranslatable words found in the queries that could not be translated by the monolingual word list due to linguistic differences between English and Zulu. The Oxford Advanced Learner's Dictionary (1995) defines linguistics as "the range of vocabulary, grammar, etc used by speakers in particular social circumstances or professional context". When investigating the improvement of proper bilingual dictionary coverage, it would help if these linguistic issues were addressed.

Table 7.1 *Untranslatable words in Zulu*

Category	Description	Occurrences	Likeliness of inclusion in dictionary
Named entities	Named entities aim to classify proper names (Elvis Presley, Microsoft), date/time (November, Tuesday, 10:30pm), measures (billion, million, rands, dollar), and other elements like email and internet addresses and phone numbers. Identification of named entities could be tricky, because of instances where the first word of a sentence is capitalised. Also, the spoken (oral) language does not indicate text with capital letters. Furthermore, case does not always indicate proper names, for instance uJames.	Due to the absence of some proper names, the average precision in retrieval was much lower. In the empirical test runs in Chapter 4 and the analysis of errors in Chapter 5, it was indicated that proper names had the highest occurrence (16%) of the dictionary-related problems, and the third highest occurrence (16%) of all problems (including translation problems).	No dictionary will include all existing proper names and other named entities, and Zulu is no exception. New proper names are constantly being created, and sometimes abbreviated to acronyms (UN/US). But, not all acronyms are proper names. Furthermore, ambiguity exists where people and months occur (April, June), as well as other categories where proper names like Precious, Prince, Presence, Innocent and Petunia occur. This is very frequent in the Zulu language.
General vocabulary	These are words that one would expect to find in the bilingual dictionary (or in the monolingual word list in this particular study).		Most of the words found in the general Zulu vocabulary are found in the dictionary.

Newly formed words		Zulu has a high occurrence of compounds (Sections 2.3.1.5 and 3.3.4.2). For instance, the word *umakhalekhukwini* is formed from *umakhala* (cry/ring) and *ekhukwini* (in the pocket).	Although both words would be found in the monolingual word list, it is highly unlikely that it would be in any dictionary. See also the criteria for borrowed words and Zululisations.
Domain-specific terminology	There are several words that are indigenous to Zulu, and that might not be in a general dictionary. Several of the paraphrased forms found in Zulu is metaphorically "made up" to illustrate the meaning.	There are several technical terms (*ikhompiyutha*) in Zulu that is not in the dictionary (yet).	These occurrences should be included in the bilingual dictionary to improve retrieval/matching results, but it might not be too realistic in the near future.
Borrowed words and Zululisations	These are words adopted from Afrikaans and English, with the same meaning; and in some instances a different spelling (computer = *ikhompiyutha*) or with a Zulu prefix added (computer = *i-computer*).	Borrowed words are very common in Zulu, with more than 48 occurrences (19%) found in the analysis of errors (see Chapter 5 for more details).	It is maybe not realistic to expect all borrowed words to be included in the dictionary though since many of the words are created "on the fly" and are not accepted as official Zulu words. One such example is *amakhemikheli*.

In light of the above categories that need to be addressed, one should determine the possibility of these words being added to the dictionary in the instance of dictionary updates. The Zulu dictionary has not been updated in thirteen years, and there might be quite a number of words that actually do deserve to be included. However, it remains to be seen when an updated dictionary becomes available.

When performing CLIR, the usual approach in managing translation ambiguity is to pass the untranslatable terms as such to the monolingual word list. However, one should not omit the accents or other diacritical signs, as this is normally an indication of tone (and important to the Zulu language). Because the focus lies on improving dictionary coverage to better manage untranslatable terms, it is proposed that all proper names and those not present in the monolingual word list be kept in their original form. In addition to the terms being written identically, the untranslatable terms would generate appropriate matches to the borrowed words. This is an important factor if cultural factors have resulted in significant language sharing over

a period. Furthermore, in the instance of English and Zulu where the languages use different writing systems, phonetic transliteration provides a useful method of achieving similar results.

The issue in improving dictionary coverage should not be the size of the dictionary. Rather, the key question should be whether you know the correct translation, and not how many translations you know.

7.4 SQ3: What can be done to contextualise language and culture in terms of CLIR?

For an English-speaking person to access a database in Zulu, the query should be constructed in English, while the English query should be directly translated into Zulu word-by-word. This Zulu query will then be run against the Zulu source database. This is where culture-related issues arise (as explained in Section 6.1). This is primarily due to the absence of an intermediary in the translation process. It is proposed that a bilingual dictionary become the intermediary. This, however, would still not solve the translation problem. Although the resulting documents will be in Zulu, it may be out of context, since the register used in the source database is directly related to the linguistic aspects of Zulu, and might not have captured all aspects related to IK.

The question now becomes: how does one capture the context of a language? It is proposed that the people that actually speak the Zulu language be included in the process. This means that (rural) translators would partly solve the intermediary problem, since context is now being captured. This is in contrast to Machine Translation, where no context exists.

Future research should investigate how the community's cooperation and involvement could be utilised–in addition to the combination of the theoretical expertise of translators and the practical life experiences of the community–to capture these culture-related issues. This would assist in the source words actually being contextualised when the category in which it exists is described by metadata. As previously stated (Section 6.3.1), the main hurdle in improving CLIR effectiveness remains with the resolving of ambiguity associated with translation. Future research should concentrate on translating adjacent words to provide the context and help with the selection of the appropriate translation. Research by Ballesteros and Croft (1997, 1998a, 1998b) describes a technique that employs co-occurrence statistics

obtained from the corpus being searched, to disambiguate dictionary-based translation. Words that are not included in phrases are translated word-by-word. However, this does not mean that they should be translated in isolation from each other. Instead, while translating a word, the other words (or their translations) form a 'context' that helps to determine the correct translation for the given word. It is proposed that this principle would work well in the Zulu-English translation process. Furthermore, Ballesteros and Croft's (1998a) assume "that the correct translations of query words tend to co-occur in target language documents and incorrect translations do not". Therefore, when provided a set of original English query words, the translator should select for each of the words the best translation word that co-occurs most often with other translation words in Zulu documents.

Although a detailed study of the Zulu linguistics (Section 3.3) offer a useful tool for analysing the structure, function and meaning of words in the Zulu language, it does not provide the necessary background to the meaning of the words. The aspect for the translation of the culture of a language also requires further investigation, especially where homonyms, paraphrased terms, register, and tone is concerned.

7.5 The road ahead

Apart from the research questions that were addressed above, future research need to determine for which language pairs it would be safe to translate not only keywords, but also entire queries. Furthermore, in terms of context, it must be determined where the cultural differences are so great that this cannot be done at all. Currently, the cultural differences can be ascribed to the fact that there is no intermediary available when English queries are translated into Zulu and back to English again. The trade of Indigenous Knowledge (IK) mostly takes place on the borders and in the border crossings between cultures where meanings and values are not codified (but rather misunderstood, misrepresented and even falsely adopted). Beyond a fixed cultural identity (that could be related to ethnicity, gender or class), so-called 'intercrossed' identities are formed by unconnected translation. This implies that former tribal societies (in this specific instance Zulu) translate their traditional 'identity' into Western forms of (for example) information technology.

One should not merely translate texts, but also have a philosophic understanding of the cultural and linguistic implications you are trying to accomplish. The world is getting larger, there are new markets, new languages and new cultures to consider as we globally provide information.

Language is more than just a means to communicate. Language is what makes us human, and aware of our surroundings. For this particular study and future research in CCIR, it becomes critical to be aware of Zulu behaviour, their beliefs and their values (stemming from personal and formal culture). Culture is the soul of the Zulu people. Through language, information can be passed on, one can learn from and connect with others, form and cultivate relationships, analyze, abstract and even evaluate facts and concepts.

To acquire an actual cross-cultural retrieval capability requires more than learning new methods of communicating information; it requires learning new ways of indexing and describing information.

Transforming one's awareness (a sense of one's personal or collective identity that includes your attitudes, beliefs and sensitivity) makes it possible to naturally accommodate cross-cultural factors when encountering people of a different culture. Instead of applying rules of "culturally-appropriate behaviour", one should rather speak and behave naturally.

Although culture does imply difference, the differences are no longer categorical; they are interactive and constantly change. This study attempted to acknowledge the cultural differences between Zulu and English, and subsequently the errors that occur when these differences are not recognised.

It is therefore suggested that interactive CLIR, which enables a user to select the best-translated keys and may add his/her own keys (in his/her personal context), might give rise to the novel concept of cross-cultural information retrieval (CCIR) (which is broader than, but encompasses CLIR). This not only concerns different languages, but also different cultures and it may bring about unique opportunities for research in CLIR (Cosijn et al., 2002a, 2002b, 2002c, 2002d).

Chapter 8

It is the quality of our work that will please God and not the quantity.

Mahatma Gandhi

8 Bibliography

ALLAN, J., CONNELL, M.E., CROFT, W.B., FENG, F-F, FISHER, D. & LI, X. 2000. InQuery and TREC-9. In: Voorhees, E. & Harman, D. (Eds.) *The Ninth Text REtrieval Conference (TREC-9)*, Gaithesburg, MD. [Online]. Available: http://trec.nist.gov/pubs/trec5/papers/umass-trec96.ps.gz [August 2001].

BALLESTEROS, L. & CROFT, W.B. 1996. Dictionary-based methods for cross-lingual information retrieval. In: Revell, N. & Tjoa, A.M. (Eds.) *Proceedings of the 7th International Database and Expert Systems Applications (DEXA) Conference on Database and Expert Systems Applications*. 9–13 September 1996. Washington: IEEE Computer Associated Press: 791–801.

BALLESTEROS, L. & CROFT, W.B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In: Belkin, N.J., Narasimhalu, A.D., Willett, P. & Hersh, W. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Philadelphia. 27–31 July 1997. New York: ACM Press: 84–91.

BALLESTEROS, L. and CROFT, W.B. 1998a. Resolving ambiguity for cross-language retrieval. In: Croft, W.B., Moffat, A., Van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. (Eds.) *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Melbourne, Australia. 24–28 August 1998. New York: ACM Press: 64–71.

BALLESTEROS, L. & CROFT, W.B. 1998b. Statistical methods for cross-language information retrieval. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval.* Boston: Kluwer Academic Publishers: 23–40.

BERGHEL, H.L. 1987. A logical framework for the correction of spelling errors in electronic documents. [Online]. Available: http://www.acm.org/~hlb/publications/logframe/logframe.html [August 2003].

BOSCH, S.E. & TALJAARD, P.C. 1988. *Handbook of IsiZulu.* Pretoria: Van Schaik.

BOSCH, S.E. & TALJAARD, P.C. 1993. *Handbook of IsiZulu.* Pretoria: Van Schaik.

BOSCH, S.E. 1999. *Two level morphological analysis: development of a morphological parser for Zulu which enables language-specific information retrieval in Zulu.* Pretoria: University of South Africa.

CALLAN, J.P., CROFT, W.B. & BROGLIO, J. 1995. TREC and TIPSTER experiments with INQUERY. *Information Processing and Management*, 31(3):327–343.

CANONICI, N.N. 1995. *Elements of Zulu morpho-syntax.* Durban: University of Natal.

CARBONELL, J., YANG, Y., FREDERKING, R., BROWN, R. D., GENG, Y., & LEE, D. 1997. Translingual information retrieval: A comparative evaluation. In: Flach, P. & Kakas, A. (Eds.) *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.* Nagoya. 23–29 August 1997.

CHISENGA, J. 2002. Indigenous knowledge: Africa's opportunity to contribute to global information content. In: Snyman, R. (Ed.) SCECSAL 2002. From Africa to the world - the globalization of indigenous knowledge systems. *Proceedings of the 15th Standing Conference of Eastern, Central and Southern African Library and Information Associations.* Caesars Gauteng Conference Centre, South Africa. 15-19 April 2002. Pretoria: LIASA.

CONOVER, W.J. 1980. *Practical non-parametric statistics.* New York: John Wiley and Sons.

COSIJN, E., JARVELIN, K., BOTHMA, T.J.D., NEL, J.G. & THEOPHANOUS, J. 2002a. Facilitating access to knowledge databases in indigenous languages. In: Retha Snyman (Ed.) SCECSAL 2002. From Africa to the world – the globalisation of indigenous knowledge systems. *Proceedings of the 15th Standing Conference of Eastern, Central and Southern African Library and Information Associations.* Caesars Gauteng Conference Centre, South Africa. 15–19 April 2002. Pretoria: LIASA: 101–117.

COSIJN, E., PIRKOLA, A., JARVELIN, K. & BOTHMA, T.J.D. 2002b. Information access in indigenous languages: a case study in Zulu. In: Fidel, R., Bruce, H., Ingwersen, P. & Vakkari, P. (Eds.) *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods.* The Information School University of Washington,

Seattle, USA. July 21–25 2002. Seattle: Libraries Unlimited: 221–238.

COSIJN, E., PIRKOLA, A., NEL, J.G., KÄNSÄLÄ, A. & BOTHMA, T.J.D. 2002c. Cross-cultural information retrieval: searching for English documents using Zulu queries. Submitted at SIGIR 2002.

COSIJN, E., BOTHMA, T.J.D., JARVELIN, K., PIRKOLA, A. & NEL, JG. 2002d. Metadata and cross-language information retrieval as complementary technologies to provide access to knowledge databases in indigenous languages. In: Bothma, T. & Kaniki, A. (Eds.) *Proceedings of ProLISSA 2002*, Pretoria, 24–25 October 2002, Pretoria: Infuse: 397–408.

COVERSON, C. 2000. *Query expansion using an interactive concept hierarchy*. [Online]. Available: http://dis.shef.ac.uk/mark/cv/publications/dissertations/coverson2000.pdf [November 2003].

DAVIS, M. 1996. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In: Voorhees, E. & Harman, D. (Eds.) *The Fifth Text REtrieval Conference (TREC-5)*, Gaithersburg, MD. 20–22 November 1996. Maryland: National Institute of Standards and Technology: 447–454.

DAVIS, M. 1998. On the effective use of large parallel corpora in cross-language text retrieval. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval*. Boston: Kluwer Academic Publishers: 11–22.

DAVIS, M. & DUNNING, T. 1995. A TREC evaluation of query translation methods for multi-lingual text retrieval. In: Harman, D. (Ed.) *The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD. 1–3 November 1995. Maryland: National Institute of Standards and Technology: 483–498.

DAVIS, M. & OGDEN, W. 1997. QUILT: Implementing a large-scale Cross-Language Text Retrieval System. In: Belkin, NJ, Narasimhalu, D. & Willett, P. (Eds.) *Proceedings of the 20th Annual International Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*. Philadelphia, PA. 27–31 July 1997. New York, NY: ACM Press: 92–98

DOKE, C.M. 1968. *Textbook of Zulu Grammar*. Cape Town: Longmans.

DOKE, C.M., MALCOLM, D.M. & SIKAKANA, J.M.A. (Eds.) 1971. *English-Zulu*

dictionary. Johannesburg: Witwatersrand University Press.

DOKE, C.M., MALCOLM, D.M., SIKAKANA, J.M.A. & VILAKAZI, B.W. (Eds.) 1990. *English-Zulu/Zulu-English dictionary*. Johannesburg: Witwatersrand University Press.

DOKE, C.M., & VILAKAZI, B.W. (Eds.) 1948. *Zulu-English dictionary*. Johannesburg: Witwatersrand University Press.

EFTHIMIADIS, E.N. 2000. Interactive query expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11): 989–1003.

ETHNOLOGUE. 2001. *Report for language code: ZUU*. [Online]. Available: http://www.ethnologue.com/show_language.asp_code=zuu. [September 2003].

FOLDOC. 1999. *Query expansion from FOLDOC*. [Online]. Available: http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi?query+expansion [September 1999].

FUJII, A. & ISHIKAWA, T. 2001. Japanese/English cross-language information retrieval: exploration of query translation and transliteration. *Computers and the Humanities*, 35(4): 389–420.

FUNG, P & MCKEOWN, K. 2003. A technical word and term translation aid using noisy parallel corpora across language groups. [Online]. Available: http://www.ee.ust.hk/~pascale/mtspecialcamera.ps [September 2003].

GACHOT, D.A., LANGE, E., & YANG, J. 1998. The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval*. Boston: Kluwer Academic Publishers: 105–118.

GLOSSARY. 2003. [Online]. Available: http://www2.plymouth.ac.uk/gateway_to_study/EngLang/glossary.htm [September 2003].

GOLLINS, T.J. 2000. *Dictionary based transitive cross-language information retrieval using lexical triangulation*. Master of Science Thesis. University of Sheffield.

GREFENSTETTE, G. (Ed.) 1998. *Cross-language information retrieval*. Boston: Kluwer Academic Press.

GREFENSTETTE, G. 1998. The problem of cross-language information retrieval. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval.* Boston: Kluwer Academic Publishers: 1–9.

GRENIER, L. 1998. *About indigenous knowledge.* [Online]. Available: http://www.nuffic.nl/ik-pages/about-ik.html [October 2002].

HANKS, P. 1983. *Collins Dictionary of the English language.* London: Collins.

HARMAN, D. 1991. How effective is suffixing? *Journal of American Society for Information Science,* 42(1): 7–15.

HASNAH, A. & EVANS, M. 1999. *Arabic/English Cross Language Information Retrieval using a bilingual dictionary.* [Online]. Available: http://www.elsnet.org/arabic2001/hasnah.pdf [November 2001].

HEDLUND,T., Pirkola, A. & Järvelin, K. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. *Information Processing and Management,* 37(1):147–161.

HILL, L. & KOCH, T. 2001. Networked organization systems: introduction to a special issue. Journal of Digital Information, 1(8). [Online]. Available: http://jodl.ecs.soton.ac.uk/Articles/v01/i08/editorial [July 2003].

HULL, D.A. 1997. Using structured queries for disambiguation in cross-language information retrieval. In: Hull, D. & Oard, D. (Eds.) *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval,* Stanford, CA. 24–26 March 1997. Stanford: AAAI Press.

HULL, D.A. & GREFENSTETTE, G. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Frei, H.P., Harman, D., Schauble, P. & Wilkinson, R. (Eds.) *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Zürich, Switzerland. 18–22 August 1996. Zurich: ACM Press: 49–57.

INDIGENOUS KNOWLEDGE AND DEVELOPMENT MONITOR. 1998. *IK Monitor (6–3) Reactions.* [Online]. Available: *http://www.nuffic.nl/ciran/ikdm/6-3/reaction.html#Three definitions* [November 2002].

INDIGENOUS LANGUAGE INSTITUTE. 2001. [Online]. Available: http://www.ipola.org [November 2001].

INTERNATIONAL INSTITUTE OF RURAL RECONSTRUCTION (IIRR). 1996. [Online]. Available: http://www.iirr.org/ [23 October 2002].

KASCHULA, R.H. & DE VRIES, I. 2000. Indigenous languages and the Internet: with LitNet as a case study. [Online]. Available: http://www.mweb.co.za/litnet/taaldebat/netlan.asp [November 2001].

KEKÄLÄINEN, J. & JÄRVELIN, K. 1998. The impact of query structure and query expansion on retrieval performance. In: Croft, W.B., Moffat, A., Van Rijsbergen, C.J., Wilkinson, R. & Zobel, J. (Eds.) *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Melbourne, Australia. 24–28 August 1998. New York, NY: ACM Press.

LANDAUER, T. & LITTMAN, M. 1991. Fully Automatic cross-language information retrieval using latent-semantic indexing. In: *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research.* Waterloo: UW Centre for the New OED: 31–38.

LITTMAN, M. L., DUMAIS, S. & LANDAUER, T.K. 1998. Automatic cross-language information retrieval using latent-semantic indexing. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval.* Boston: Kluwer Academic Publishers: 51–62.

LOPRESTI, D. & WILFONG, G. 1999. Cross-Domain Approximate String Matching. In: Baeza-Yates, R., Ribeiro, B. & Ziviani, N. (Eds.) *The 6th International Symposium on String Processing and Information Retrieval,* Cancun, Mexico. [Online]. Available: http://bell-labs.com/org/1133/Research/Document Analysis/papers/xdomain.pdf [April 2002].

MATEEV, B., MUNTEANU, E., SHERIDAN, P., WECHSLER, M. & SCHAUBLE, P. 1996. ETH TREC-6: Routing. Chinese, Cross-Language and Spoken Document Retrieval. In: Voorhees, E.M. & Harman, D.K. (Eds.) *Proceedings of the 6th Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST),* Gaithersburg, MD. 19–21 November 1997. Maryland: National Institute of Standards and Technology: 623–636.

MAXWELL, M. 1997. *Computational Linguistics.* [Online]. Available: http://www.sil.org/computing/hermitcrab/ [November 2002].

MENG, H., CHEN, B., GRAMS, E., KHUDANPUR, S., LO, W-K., LEVOW, G-A, OARD, D., SCHONE, B., TANG, K., WANG, H-M., & WANG, J.Q., 2000. Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval. In: Allan, J. (Ed.) HLT 2001, *Human Language Technology Confererence*, 18–21 March 2001, San Diego, California.

MILSTEAD, J. & FELDMAN, S. 1999. *Metadata: Cataloguing by any other name.* [Online]. Available: http://www.infotoday.com/online/OL1999/milstead1.html [September 2003].

NGULUBE, P. 2002. Strategies for managing and preserving indigenous knowledge in the knowledge management era. In: Snyman, R. (Ed.) SCECSAL 2002. From Africa to the world – the globalisation of indigenous knowledge systems. *Proceedings of the 15th Standing Conference of Eastern, Central and Southern African Library and Information Associations*. Caesars Gauteng Conference Centre, South Africa. 15–19 April 2002. Pretoria: LIASA: 61–69.

OARD, D. 1997. *Serving users in many languages.* [Online]. Available: http://www.dlib.org/dlib/december97/oard/12oard.html [August 2003].

OARD, D. 2001. *Cross-language information retrieval defined.* [Online]. Available: http://raven.umd.edu/dlrg/clir/mlir_definition.html [November 2002].

OARD, D. & DIEKEMA, A. 1998. Cross-Language Information Retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33: 223–256.

OXFORD ADVANCED LEARNER'S DICTIONARY. 1995. Oxford University Press.

PETERS, C. & PICCHI, E. 1997. Using linguistic tools and resources in cross-language information retrieval. In: *Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.* Stanford, CA. [Online]. Available: http://www.ee.umd.edu/medlab/filter/sss/papers [November 2002].

PFEIFER, U., POERSCH, T. & FUHR, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing and Management*, 32(6): 667–679.

PIRKOLA, A. 1998. The effects of query structure and dictionary set-ups in dictionary-based cross-language information retrieval. In: Wilkinson, R., Croft, B. and van Rijsbergen, C. (Eds.) *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Melbourne, Australia. 24–28 August 1998. New York, NY: ACM Press: 55–63.

PIRKOLA, A. 1999. *Studies on linguistic problems and methods in text retrieval: the effects of anaphor and ellipsis resolution in proximity searching and translation and query structuring methods in cross-language retrieval.* PhD Dissertation, University of Tampere, Department of Information Studies, Acta Universitatis Tamperensis 672.

PIRKOLA, A., HEDLUND, T., KESKUSTALO, H., & JÄRVELIN, K. 2001. Dictionary-based cross-language information retrieval: problems, methods and research findings. *Information Research*, 4(3/4): 209–230.

PIRKOLA, A., KESKUSTALO, H., LEPPÄNEN, E., KÄNSÄLÄ, A., JÄRVELIN, K. 2002a. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2) [Online]. Available: http://InformationR.net/ir/7-2/paper126.html [September 2002].

PIRKOLA, A., LEPPÄNEN, E. & JÄRVELIN, K. 2002b. The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. *Information Research*, 7(2). [Online]. Available: http://InformationR.net/ir/7-2/infres72.html [September 2002].

PORTER, M.F. 1980. An algorithm for suffix stripping. *Program*, 14: 130–137.

REHDER, B., LITTMAN, M.L., DUMAIS, S. & LANDAUER, T.K. 1998. Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In: Voorhees, E. & Harman, D. (Eds.) *Proceedings of the Sixth Text Retrieval Conference (TREC6), National Institute of Standards and Technology (NIST).* Gaithersburg, MD. 19–21 November 1995. Maryland: National Institute of Standards and Technology: 233–240.

ROBERTSON, A.M. & WILLETT, P. 1998. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1): 48–69.

RUIZ, M.E. & SRINIVASAN, P. 1998. Cross-language information retrieval: an analysis of errors. [Online]. Available: http://www.cs.uiowa.edu/~mruiz/papers/asis96.ps.gz [August 2001].

SALTON, G. 1989. *Automatic text processing: the transformation analysis and retrieval of information by computer.* Reading, MA: Addison-Wesley.

SELBERG, E. 1997. *Query expansion and relevance feedback.* [Online]. Available: http://selberg.org/homes/speed/papers/generals/generals/node4.html [November 2003].

SHERIDAN, P. & BALLERINI, J.P. 1996. Experiments in Multilingual Information Retrieval using the SPIDER System. In: Frei, H.P., Harman, D., Schauble, P. & Wilkinson, R. (Eds.) *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland. 18–22 August 1996. Zurich: ACM Press: 58–65.

SHERIDAN, P. & SHÄUBLE, P. 1997. Cross-Language Multi-Media Information Retrieval. In: *Proceedings of the Third DELOS workshop, Cross-Language Information Retrieval, ERCIM Workshop*. Zurich, Switzerland. 5–7 March 1997.

SHERIDAN, P., BRASCHLER, M. & SCHÄUBLE, P. 1997. Cross-Language Information Retrieval in a Multilingual Legal Domain. In: Peters. C., and Thanos, C. (Eds.) *Research and Advanced Technology for Digital Libraries. First European Conference, ECDL '97*, Pisa, Italy, 1–3 September 1997, pp. 253–268.

SPERER, R. & OARD, D. 2000. Structured translation for cross-language IR. In: Belkin, N.J., Ingwersen, P. & Leong, M-K. (Eds.) *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece. 24–28 July 2000,pp. 120–127.

STATISTICS SOUTH AFRICA. 2003. [Online]. Available: http://www.statssa.gov.za [August 2003].

THE INFORMATION RETRIEVAL GROUP. 2002. [Online]. Available: http://www-2.cs.cmu.edu/People/callan/IRGroup/ [November 2002].

UCLA LANGUAGE MATERIALS PROJECT. 2001. *Zulu profile.* [Online]. Available: http://www.lmp.ucla.edu/profiles/profz01.htm [January 2002].

WASSERMAN, H. 2000. *Between the local and the global: South African languages and the Internet.* [Online]. Available: http://www.mweb.co.za/litnet/seminarroom/Hwasserman.asp [November 2001].

WILKES, A. 1988. *IsiZulu sezikhuthuli.* Pretoria: De Jager-Haum.

WILKES, A. 2001. *Class Notes: IsiZulu 201*, Pretoria: University of Pretoria.

WILKES, A. & NKOSI, N. 1995. *Zulu: a complete beginner's course*. Johannesburg: Hodder and Stoughton.

WORLD BANK GROUP. 1998. *What is IK?* [Online]. Available: http://www.worldbank.org/afr/ik/ what.htm [November 2002].

YAMABANA, K., MURAKI, K., DOI, S. & KAMEI, S. 1998. A language conversion front-end for cross-language information retrieval. In: Grefenstette, G. (Ed.) *Cross-Language Information Retrieval*. Boston: Kluwer Academic Publishers: 93–104.

ZOBEL, J. & DART, P. 1995. Finding approximate matches in large lexicons. *Software–practice and experience*, 25(3): 331–345.

Appendix A

Table A1.1 CLEF Query C041 Translations

	Query	Translation	Description	Translation
Original English CLEF query	Pesticides in baby food		Find reports on pesticides in baby food	
Official translation	Amakhemikheli abulala zonke izifo ezinengozi ekudleni kwezingane	Chemicals that put to death all diseases that is dangerous in the food of small children.	Thola imibiko ephathelene namakhemikheli abulala zonke izifo ezinengozi ekudleni kwezingane.	Find reports that mention chemicals that kill all diseases in the food of small children.
Mother tongue translation	Umuthi wokubulala izinambuzane edudleni kwabantwana.	Medicine that kills insects in the food of babies.	Thola imininingwane ngemithi yokubulala izinambuzane edudleni kwabantwana.	Find detail about medicine that kills insects in the food of babies

Table A1.2 CLEF Query C041 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
abulala	1	bulala	lying position recumbent posture	
	2	**bulala**	**kill murder destroy harm hurt injure poisonous**	
	3	-nyabulala isinyabulala izinyabulala	very young child person weak illness	
amakhemikheli	1	kheli ikheli amakheli	address mannerism private character	No match – amakhemikheli is a borrowed word adapted to Zulu pronunciation.
	2	-kheji ikheji amakheji	cage bird cage transport cage	
	3	-kheli umkheli abakheli	person addresses letter	
ekudleni	1	-nkulukundleni unkulukundleni onkulukundleni	species caterpillar species encased grubs	No match – ekudleni is formed by uku+dla+ini. By means of vowel coalescence the -dla+ini becomes -dleni.
	2	-nkulukundleni unkulukundleni onkulukundleni	species caterpillar species encased grubs	

	3	ekuqaleni	beginning	
ephathelene	1	phathelela	grip tightly hold lay hands make constant reference	
	2	**phathela**	**handle carry treat mention**	
	3	phathelana	concerned connected relate	
ezinengozi	1	-nengo isinengo izinengo	disgusting thing nauseating object	
	2	-nengenenge unengenenge izinengenenge	endless drawn matter continuous following persecution	
	3	**-ngozi ingozi izingozi**	**severe wound throbbing pain wound accident misfortune danger risk**	
imibiko	1	**-biko umbiko imibiko**	**report message announcement**	
	2	-boko umboko imiboko	soft glandular swelling	
	3	-bino umbino imibino	obscene language	
izifo	1	**-fo isifo izifo**	**disease sickness death**	
	2	-foco isifoco izifoco	dent pitting	
	3	foko isifoko izifoko	native womans top knot	
kwezingane	1	**-ngane ingane izingane**	**infant small child foetus**	
	2	-ngane isingane izingane	sweetheart close friend	
	3	-nganeko inganeko izinganeko		
thola	1	**thola**	**obtain find**	

	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcety scantiness	
zonke	1	**zonke**	Inclusive quantifier Class 10	
	2	onke	Inclusive quantifier Class 6	
	3	konke	Inclusive quantifier Class 15	

Table A2.1 CLEF Query C042 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	UN/US Invasion of Haiti		Find documents on the invasion of Haiti by U.N./US soldiers	
Official translation	Ukuhlaselwa kwezwe laseHaite iNhlangano yeZizwe neMelika.	*The invasion/attack of the country of Haiti by the United Nations and America.*	Thola imibhalo ngokuhlaselwa kwezwe laseHaiti ngamasosha weNhlangano yeZizwe neMelika.	Find documents about the invasion of the country of Haiti by soldiers/troops of the United Nations and America.
Mother tongue translation	Ukuhlaselwa kweHaiti yi-UN/US.	The invasion of Haiti by UN/US.	Thola imininingwane ekuhlaselweni kweHaiti amabutho ase-UN/US.	Find details about the invasion/attack of the country of Haiti by soldiers/troops of the UN/US.

Table A2.2 CLEF Query C042 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Haiti	1	-nayiti inayiti izinayiti	needle	No match – Proper name
	2	-titi utiti	porridge mashed sweet potatoes pumpkins	
	3	-nayiti inayiti izinayiti	needle	

Melika	1	**-Melika iMelika amaMelika**	**American person Congregational Church**	Although also a proper name as in the case of Haiti, the word was found in the dictionary.
	2	-Melika iMelika	American Board Missions Congregational Church	
	3	-Melika iMelika amaMelika	American person Congregational Church	
enhlangano	1	**-hlangano inhlangano izinhlangano**	**meeting assembly point meeting confluence junction meeting ways place assembly league agreement alliance pact compact**	A match was found, but this word is also classified as a homonym (same word for different meanings "meeting", "junction", "league" and "alliance").
	2	-hlangano umhlangano imihlangano	meeting assembly people	
	3	-hlangano inhlangano izinhlangano	meeting assembly point meeting confluence junction meeting ways place assembly league agreement alliance pact compact	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Another example of a homonym – writing VS large blanket.
	2	**-balo umbalo imibalo**	**writing document mark entry scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
kwezwe	1	-zwezwe umzwezwe imizwezwe	spur cock	No match – kwezwe is formed by kwa+izwe through vowel coalescence.
	2	-zwezwe umzwezwe imizwezwe	spur cock	
	3	-nkweza inkweza izinkweza	day	

ngamasosha	1	**-sosha isosha amasosha**	**soldier**	An example of a borrowed word.
	2	-sosa isosa amasosa	saucer	
	3	-solosha isolosha amasolosha	hoof horse ox	
ngokuhlaselwa	1	ngokuhlwa	nightfall dusk	
	2	**hlasela**	**war attack invade hunt**	
	3	-hlaseli umhlaseli abahlaseli	attacker invader hunter	
thola	1	**thola**	**obtain pick come adopt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
yezizwe	1	**-zwe isizwe izizwe**	**nation tribe clan state rapidly spreading brain disease mania**	Although a match was made, this word is a good example of a homonym, since the word could easily have been mistranslated to "rapidly spreading brain disease".
	2	-yezi isiyezi iziyezi	dizziness giddiness vertigo	
	3	-yeziyezi amayeziyezi	mucus passed faeces stool inconsistencies things readily mix	

Table A3.1 CLEF Query C043 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***El Nino and the Weather***		Find reports explaining the "El Nino" phenomenom and its repercussions on the world's weather (including effects on temperature, air pressure, rain fall, etc.).	
Official translation	El Nino nesimo sezulu	El Nino and the shape/form of the sky	Thola imibiko nge-El Nino nemiphumela yayo esimeni sezulu umhlaba wonke (kanye nemiphumela yalokho kwizinga lokushisa, isisindo somoya, nokunetha kwezulu njll.).	Find reports of El Nino and the after effects of it in the weather of the whole world together with the after effect of it on the process of becoming hot, the influence of the wind and the coming of wetness to the sky.
Mother tongue translation	I-El Nino nesimo sezulu	El Nino and the form/shape of the sky.	Thola imininingwane echaza ngezigigaba ze-El Nino nemiphumela yazo kwisimo sezulu somhlaba wonke (ufake nemiphumela yokushisa, nemimoya, nemvula).	Find details that explain the characteristics of El Nino and the after effects of it on the shape of the weather of the world that is whole (and give the after effects of the temperature, the air pressure and the rainfall).

Table A3.2 CLEF Query C031 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
El	1	-thongo ithongo amathongo	ancestral spirit	No match, since El is part of El Nino – a proper name.
	2	babela	burn grass	
	3	babela	pungent bitter acrid burning desire consistently aimed	
Nino	1	-fino isifino izifino	daily ordinary food kind tasty dish European food	No match, since Nino is part of El Nino – a proper name not found in the dictionary.

	2	-fino umfino imfino imifino	ordinary food green vegetable edible herb cooked food	
	3	-sino umsino imisino	dance dancing party	
esimeni	1	-sime umsime imisime	crutch appliance aid walking crossing	No match – esimeni is formed by isimo+ini through adding a locative form (-ini).
	2	-eno umeno imeno	thick tangled overgrowth grass bush creepers jungle	
	3	-menywa isimenywa izimenywa	guest invited person	
imibiko	1	**-biko umbiko imibiko**	**report message announcement**	
	2	-boko umboko imiboko	elephants trunk penis horse proboscis insect	
	3	-bino umbino imibino	obscene language	
isisindo	1	**-sondo isisindo**	**weight measure heaviness dignity influence personality importance**	A match was found, but it must be noted that the word forms part of a paraphrase "isisindo somoya" – the influence of the wind < air pressure.
	2	-sindo isindo amasindo	rescue escape	
	3	-sindo umsindo imisindo	noise din uproar quarrel	
kanye	1	kanye	way	There is no match, since the word is an enumerative formed by a+nye.
	2	kanyekanye	time	
	3	-nkanyezi inkanyezi izinkanyezi	star firefly	
kwezulu	1	**-zulu izulu**	**sky heaven weather lightning**	
	2	-zulu isizulu	Zulu language mannerisms	

	3	-zulu ubuzulu	membership Zulu race intoxicating drink made juice palm	
kwizinga	1	-nga- inga- isinga- izinga-	portion	No match – part of a paraphrase that means "the process of becoming hot" > temperature
	2	-nga isinga izinga	direction near	
	3	-zinga izinga amazinga	ridge corrugation running elevation two closely placed grooves ridge two valleys annular ring tree corrugation ring horn scale snake fish tortoise millipede	
lokushisa	1	-shisa ushisa oshisa	native eating house	No match – part of paraphrase (kwizinga lokushisa) that means "the process of becoming hot" > temperature
	2	-moyushisa umoyushisa	medicine distilled medicated vapour	
	3	shushisa	make restless urge encourage invigorate	
nemiphumela	1	**-phumela umphumela imiphumela**	**small shrub love charm effect**	Indeed matching "effect", but also homonym that could mean "small shrub" or "love charm".
	2	phumela	come rise leave discharge go relieve nature	
	3	-nomphumela unomphumela onomphumela	species bushveld shrub bearing fruit	
nesimo	1	-nesi inesi amanesi	nurse	No match because of vowel coalescence – na+isimo = nesimo
	2	-nesi unesi onesi	nurse hospital nurse	
	3	-nesi ubunesi	nursing	
nge	1	nge	first person singular	No match – nge in this context takes the meaning of an object concord in Class 1 (first person singular.)
	2	-nge-	negative verb	
	3	-nga- inga- isinga- izinga-	place overgrown mimosa trees tall species fever mimosa	

njll	1	njalo	always continually	No match because of njll being an enclitic.
	2	-njalo	like that such as that	
	3	-njalo injalo izinjalo	seed potato seed tuber	
nokunetha	1	**netha**	**wet soaked rain leak rain sluggish inactive drowsy soft brained**	Homonym. Nokunetha also formed through vowel coalescence.
	2	-netha inetha izinetha	rain	
	3	-netha inetha amanetha	net	
sezulu	1	**-zulu izulu**	**sky heaven weather lightning**	
	2	-zulu isizulu	Zulu language mannerisms	
	3	-zulu ubuzulu	membership Zulu race intoxicating drink made juice palm	
somoya	1	**-moya ummoya umoya imimoya immoya**	**wind air breath spirit soul life rumour nonsense**	A good example of a homonym. Somoya also part of paraphrased term – "the influence of the wind"
	2	-oya umoya	spirit	
	3	-momoyi imomoyi amamomoyi	handsome person full fresh healthy cheeks	
thola	1	**thola**	**obtain find come adopt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
umhlaba	1	-hlaba umhlaba imihlaba	several species large aloe	

	2	**-hlaba umhlaba imihlaba**	**earth world land country something value**	
	3	-hlababa umhlababa	species thorny shrub	
wonke	1	**wonke**	**Inclusive quantifier Class 1**	
	2	-wonke uwonke	public masses	
	3	onke	Inclusive quantifier Class 6.	
yalokho	1	lokho	forthwith	No match, since the word is formed by means of the possessive concord (ya-) + the second position demonstrative (-lokho).
	2	ngalokho	account concerning	
	3	kalokhu	moment	

Table A4.1 CLEF Query C044 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	Indurain wins Tour		Reactions to the fourth Tour de France won by Miguel Indurain	
Official translation	U-Indurain uwina iTour de France.	Indurain wins the Tour de France.	Ukuphatheka kwabantu ngeTour de France yesine ezuzwe nguMiguel Indurain.	To deal with the feelings of the people of the Tour de France that is fourth which was won by Miguel Indurain.
Mother tongue translation	U-Indurain uphumelela emncintiswaneni wohambo.	Indurain succeeded in the competition of the journey that is endless.	Imiphumela yomuncintiswano i-"Tour de France" wesine oklonyeliswe u-Miguel Indurain.	How have the people taken up the after effects of the success of the competition of the Tour de France that is fourth, which was given the prize/reward by Miguel Indurain

Table A4.2 CLEF Query C044 Matching

Word in official translation		Dictionary entry	Translations	Comment
France	1	-ncengancenga incengancenga izincengancenga	small fretful peevish child species rush mat making	No match, since France is a proper name. Interesting though that there is a dictionary entry under eFulansi, which is a Zululised word.
	2	nce	tinkling metallic sound ticking	
	3	-nce isince izince	incomplete shortened object stump object portion short haftless spear blade waning moon	
Indurain	1	-duva induva izinduva	thing disregarded cared little value	Another proper name.
	2	-duna induna izinduna	male animal officer state army captain overseer head man councillor	
	3	-duma induma	species rush mat making	
Miguel	1	-gele igele amagele	species veld plant edible bulbous roots	No match because of the proper name Miguel.
	2	-gubela igubela amagubela	plume long feather	
	3	-gwele igwele amagwele	species succulent rock plant pink flowers	
Tour	1	wou	pleasurable excitement	No match, although "tour" is not a proper name. In this case the word is paraphrased to "the journey that is endless" in the mother tongue, and to iTour (borrowed word) in the dictionary.
	2	-jeri ujeri ojeri	Jack fruit tree	
	3	-Thuthuthu iThuthuthu	Cetshwayos regiments	
de	1	-de	act	The word is not matched to a correct dictionary entry, since it forms part of a phrase – Tour de France. In this context the word does not have a meaning.
	2	-de ubude	length height depth distance	
	3	-nde umunde iminde	species flying ant	
ezuzwe	1	-zwela uzwela	sensitiveness touchiness tender feeling	No match because of vowel coalescence (a+i = e), and -zuzwa becoming -zuzwe.

	2	-zwezwe umzwezwe imizwezwe	spur cock	
	3	-zwenda uzwenda izinzwenda	tough natured object	
kwabantu	1	**-ntu umuntu abantu**	**person man human being member African Native race black man person human feelings subject servant dependant**	A match was made, but because the word is a homonym, the meaning could be mistranslated to "person" or "race" or "human feelings".
	2	-ntakwabo umntakwabo abantakwabo	member mothers hut blood relation sweetheart	
	3	kwalahlabantu	place execution	
ukuphatheka	1	phaphatheka	lose head excited act speak nervously rush away run wildly stampede fade lose colour	No match, due a the word being formed from uku+phatha+eka. Vowel elision occurs (-phatha- becomes -phath-) and a verbal extension is added.
	2	-phaphatheka iphaphatheka amaphaphatheka	nervous excitable person timid person animal	
	3	-phaphatheka umphaphatheka imiphaphatheka	faded discoloured object	
uwina	1	**wina**	**win successful**	
	2	ncwina	sing sweetly twitter complain show discontent whine	
	3	-mina umina izimina	challenger	
yesine	1	-yesi iyesi amayesi	light refreshment	The word was not matched because of vowel coalescence (ya+i=ye), and the stem being -ne (four).
	2	-lwesine ulwesine	Thursday	
	3	-yesi iyesi amayesi	light refreshment	

Table A5.1 CLEF Query C046 Translations

	Query	Translation	Description	Translation
Original English CLEF query	Embargo on Iraq		***What effect has the U.N. embargo had on the lives of the Iraqi people?***	
Official translation	Ukuvinjelwa kohwebo e-Iraq.	Preventing trade in Iraq.	Kube miphumela mini ukuvinjelwa kohwebo kwiNnhlangano yeZizwe ezimpilweni zabantu base-Iraq?	Of what kind were the after effects to prevent trade of the United Nations in the lives of the people of Iraq?
Mother tongue translation	Ukuvalelwa kokushintshisana ngempahla e-Iraqi.	The closing of the mutual exchange of goods/things in Iraq.	Ibe namiphi imiphumela ezimpilweni zabantu base-Iraqi, I-UN ngenkathi ivalela ukushintshisana kwempahla	Which were the after affects on the lives of the people of Iraq, by the UN by the time it was closed the mutual exchange of goods.

Table A5.2 CLEF Query C046 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Iraq	1	-aqu isaqu izaqu	hunting song	The word was not matched because of Iraq being a proper name. In Zulu only the class prefix is added to the proper name, in this case the -i(li) from Class 5.
	2	qaqa	rip rip open undo seam explain difficulty surround	
	3	-qaqa amaqaqa	eyebrow bones protruding	
ezimpilweni	1	-phila impila izimpila	species veld herb tonic girls early stage menstruation	The stem in this case is -pilo. No match because of the forming of a locative (o > weni).
	2	-mpi impi izimpi	regiment army military force encounter fight engagement battle war hostile person band foe enemy	
	3	-mpimpiliza impimpiliza izimpimpiliza	spinning top	
kohwebo	1	**-hwebo umhwebo imihwebo**	**trade barter**	

	2	-khwebu isikhwebu izikhwebu	mealie cob childless wife husband lonely object	
	3	khweba	gather crops matured	
kube	1	kube ngakube ngaye-kube ngabe-kube ngasekube ngahle-kube		The word has no meaning, and is in a conjuntive form.
	2	**ukube**	**conjunction constructions**	
	3	ukuba	conjunction constructions	
kwenhlangano	1	**-hlangano inhlangano izinhlangano**	**meeting assembly confluence junction meeting ways league agreement alliance pact compact**	This word is part of a paraphrase "the nations that is united", but matches on its own as well to "alliance".
	2	-hlangano umhlangano imihlangano	meeting assembly people	
	3	hlangana	come meet assemble meet come come join unite close contact thick dense	
mini	1	mini	day	No match was possible because the word is formed from the interrogative -ni, meaning "of what kind".
	2	-mini imini izimini iMini	day daytime midday noon	
	3	emini	day during day	
miphumela	1	**-phumela umphumela imiphumela eng eng**	**small shrub love charm effect**	Although a match was made, it could have easily been mistranslated to "love charm" or "small shrub" because the word is a homonym.
	2	phumela	come eve rise discharge relieve nature	
	3	-xhumela ixhumela amaxhumela	high heel shoe	

uhwebo	1	**-hwebo umhwebo imihwebo**	**trade barter**	
	2	-hwebi umhwebi abahwebi	trader merchant	
	3	-gwebo isigwebo izigwebo	decision judgement hooked instrument	
ukuvimbela	1	**vimbela**	**close way prevent block way**	
	2	-vimbela ivimbela amavimbela	preventive charm medicine fabulous water snake	
	3	-vimbelo isivimbelo izivimbelo	obstacle	
yezizwe	1	**-zwe isizwe izizwe**	**nation tribe clan state rapidly spreading brain disease mania**	This is also a homonym that could have been mistranslated to "rapidly spreading brain disease". In this case the word forms part of the paraphrase "the nations that is united".
	2	-yezi isiyezi iziyezi	dizziness giddiness vertigo	
	3	-yeziyezi amayeziyezi	mucus passed feaces stool	
zabantu	1	**-ntu umuntu abantu**	**man person human member African Native race black man person human feelings subject dependant servant**	
	2	bantu	people wonderful believe say dear	
	3	-Ndabazabantu uNdabazabantu oNdabazabantu	Native Affairs Department Ministry Native Affairs Native Commissioner Minister Native Affairs	

Table A6.1 CLEF Query C047 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Russian Intervention in Chechnya***		What are the reasons for the military intervention of Russia in Chechnya?	
Official translation	Ukungena kweRussia eChechnya.	The entering/invasion of Russia in Chechnya.	Kwakungazizathu zini ukungena kwezombutho kweRussia eChechnya?	The entering of the soldiers by Russia in Chechnya was for what reasons?
Mother tongue translation	Ukuhlaselwa kweRussia eChechnya.	The invasion of Russia in Chechnya.	Iziphi izizathu zokuhlasela kweRussia eChechnya?	Which are the reasons for the invasion of Russia in Chechnya?

Table A6.2 CLEF Query C047 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Chechnya	1	-Chenyane iChenyane amaChenyane	regiment girls formed Dingane	There is no match for this word, since it is a proper name and only takes a class prefix.
	2	chiya	stay support prop	
	3	chaya	slice strips tear shreds spread lay flat make long body incisions	
Russia	1	-susi umsusi abasusi	starts originates remover	There is no match for this word, since it is a proper name and only takes a class prefix.
	2	-busisi umbusisi ababusisi	bestows blessings	
	3	-si usi	smell	
kwakungazizathu	1	**-zathu isizathu izizathu**	**reason excuse cause**	
	2	azi ukuzazi ukungazi	conceited pay particular attention personal appearance rank ignorance	
	3	-thungakazana umthungakazana imithungakazana	species forest tree	

kwezombutho	1	-butho umbutho imibutho	gathering meeting study worship conference	
	2	**-butho ibutho amabutho**	**regiment age grade band enrolled Zulu king member regiment warrior soldier members age grade**	
	3	-butho umbutho imibutho	gathering meeting study worship conference	
ukungena	1	**ngena**	**enter come invade take place cow milking purposes take wife widow deceased brother enter new year season know intimately**	Another excellent example of a homonym that could have been mistranslated to "milking of a cow" or "the purpose of taking a wife", "to enter the new year" or "to know intimately".
	2	-nene ukunene	right side	
	3	-ngeno ubungeno	way approach entrance admission	
zini	1	-nini isinini izinini	maize store	No match was possible because the word is formed from the interrogative -ni, meaning "of what kind".
	2	-ini	locative suffix forming locative adverbs	
	3	-nina isinina izinina	central part side hut wattles arched	

Table A7.1 CLEF Query C048 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Peace-Keeping Forces in Bosnia***		Reasons for the withdrawal of United Nations (UN) peace-keeping forces from Bosnia	
Official translation	Amabutho okukhipha uxolo eBosnia.	Soldiers that take peace in Bosnia.	Izizathu zokuphuma kwamabutho oxolo weNhlangano yeZizwe eBosnia.	Reasons for the exit of soldiers of peace of the United Nations from Bosnia.
Mother tongue translation	Amabutho okugcina ukuthula eBosnia	Soldiers that guard/keep peace in Bosnia.	Izizathu zokukhishwa kwamabutho agcina ukuthula eNhlangano yeZizwe ezihlangene eBosnia.	Reasons for the extraction/pull-out of soldiers/troops of peacekeeping of the United Nations from Bosnia.

Table A7.2 CLEF Query C048 Matching

Word in official translation		Dictionary entry	Translations	Comment
Bosnia	1	-nina unina onina	mother foster mother mistress parent plant shoots grow	There is no match for this word, since it is a proper name and only takes a class prefix.
	2	-nonina unonina ononina	woman variety rank status	
	3	-sosipani isosipani amasosipani	saucepan	
amabutho	1	**-butho ibutho amabutho**	**regiment age grade band enrolled Zulu king member regiment warrior soldier members age grade**	
	2	-butha ibutha amabutha	love charm emetic	
	3	-qalabutho ingqalabutho izingqalabutho	first born founding member club people age grade	
izizathu	1	**-zathu isizathu izizathu**	**reason excuse cause**	
	2	-zathu isizathu	sharp tongue	
	3	-thathu isithathu izithathu	third place third	
kwenhlangano	1	**-hlangano inhlangano izinhlangano**	**meeting assembly confluence junction meeting ways league agreement alliance pact compact**	This word is part of the paraphrase United Nations.
	2	-hlangano umhlangano imihlangano eng	meeting assembly people	

	3	hlangana	come meet assemble meet come come join unite close contact thick dense	
okukhipha	1	**khipha**	**take exclude turn expel extract draw purge**	
	2	-khiphanoni ukhiphanoni okhiphanoni	species hairy shrub	
	3	phipha	wipe away dirt remove ordure wipe child stool help difficulty assist helpless	
oxolo	1	-xolo ixolo amaxolo	bark tree species small veld plant	
	2	**-xolo uxolo**	**peace quiet calm goodwill pardon forgiveness**	
	3	-gxolo igxolo amagxolo	bark tree species small veld plant	
uxolo	1	-xolo ixolo amaxolo	bark tree species small veld plant	
	2	**-xolo uxolo**	**peace quiet calm goodwill pardon forgiveness**	
	3	-gxolo igxolo amagxolo	bark tree species small veld plant	
yezizwe	1	**-zwe isizwe izizwe**	**nation tribe clan state rapidly spreading brain disease mania**	This word is part of the paraphrase United Nations.
	2	-yezi isiyezi iziyezi	dizziness giddiness vertigo	
	3	-yeziyezi amayeziyezi	mucus passed faeces stool	
zamabutho	1	**-butho ibutho amabutho**	**regiment age grade band enrolled Zulu king member regiment warrior soldier members age grade**	
	2	-butha ibutha amabutha	love charm emetic	
	3	-qalabutho ingqalabutho izingqalabutho	first born founding member club people age grade	
zokuphuma	1	phuphuma	overflow bubble overcome emotion miscarriage	The word is not matched to any dictionary entry, since the

	2	-phuphuma iphuphuma amaphuphuma	species tree love charm	word is formed by za+uku+phuma. Through vowel coalescence (a+u=o) za+uku becomes zoku-
	3	-phuphuma impuphuma izimpuphuma	overflow superfluity	

Table A8.1 CLEF Query C051 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***World Soccer Championship***		Find documents reporting on the final game of the world soccer championship of 1994.	
Official translation	Izingcweti zebhola lomhlaba.	Experts of the ball of the world.	Thola imibhalo ebika ngomdlalo wokugcina wezingcweti zebhola lomhlaba ngo-1994	Find documents that report on the game that is final of the experts of the ball of the world that was in 1994
Mother tongue translation	Imidlalo yompetha bomhlaba kwibhola likanobhuduzwayo.	The game of "cunning people" (champions) of the world of the ball in soccer.	Thola amaphepha abika ngomdlalo wokugcina wompetha bomhlaba kubhola likanobhuduzwayo onyakeni ka-1994	Find papers that report on the game that is final of the "cunning people" (champions) of the world of the ball of soccer of the year that is 1994.

Table A8.2 CLEF Query C051 Matching

Word in official translation		Dictionary entry	Translations	Comment
bango-1994	1	-bango umbango imibango	charm medicine incantation calculated bring family strife instrument achieve witchcraft	There is no match in this case, since the numeral is preceded by an inflected prefix.
	2	-bango isibango izibango	family tribal dispute property heirship feud vendetta contention girl	
	3	-cabango ucabango	xiphisternum cartilage end breastbone spot below solar plexus	

bebhola	1	hola	draw drag haul pull pull rope absorb draw smoke pipe inhale draw water pipe suck lead water lead entice guide draw pay	There is no match for this word, since the stem is -bhola.
	2	shola	pick	
	3	xhola	inquire closely investigate	
bomhlaba	1	**-hlaba umhlaba imihlaba**	**earth world land country no value**	
	2	hlaba	pierce stab stick gore prick thrust kill slaughter attack pain cause sharp stabbing pain reach do perfectly hurt feelings wound mentally criticize review drive	
	3	-hlaba ihlaba amahlaba iHlaba	sharp pain prickly aloe dressing hides sonchus dregeanus milk thistle strong pungent thing aloes bitter beer soil grave witchcraft cause lung disease mixed food regiment Dingane	
ebika	1	**bika -vika**	**report introduce warn omen**	Another good example of a homonym, since -bika could have been mistranslated to "omen" instead of "report".
	2	-bika ibika amabika	harbinger omen term applied bird insect species ant white spots introduction	
	3	-bibika imbibika izimbibika	cry baby mournful faced person	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Another homonym that could have mistranslated to "large blanket".
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
ngomdlalo	1	**-dlalo umdlalo imidlalo**	**game sport amusement entertainment concert**	
	2	-dlalo isidlalo izidlalo	plaything laughing stock place lightning given strike pneumonia	

	3	-gomela ingomela izingomela	asseveration assertion determination	
ongcweti	1	-ngcweti ubungcweti	expertness experience faculty	This is an interesting match, since champion is paraphrased to "experts of the ball of the world" – which is exactly what an champion is!
	2	**-ngcweti ingcweti izingcweti**	**expert gifted experienced**	
	3	ngcwetiza	make clever give experience	
thola	1	**thola**	**obtain pick come find adopt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
wokugcina	1	**gcina**	**terminate come end keep preserve take care observe make stick fast firm**	The word is part of a paraphrase for "final game".
	2	-gcina ingcina izingcina	securing making fast species lily glutinous substance	
	3	ekugcineni	completion	
wongcweti	1	-ngcweti ubungcweti	expertness experience faculty	
	2	**-ngcweti ingcweti izingcweti**	**expert gifted experienced**	
	3	ngcwetiza	make clever give experience	

Table A9.1 CLEF Query C052 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Chinese Currency Devaluation***		Find documents describing the reasons and effects of the devaluation of Chinese currency.	
Official translation	Ukwehla kwezinga lomnotho eChina.	The decline of the ridge/curve of the riches (economy) of China.	Thola imibhalo echaza izizathu nemiphumela yokwehla kwezinga lomnotho eChina.	Find documents that explain the reasons and after effects of the decline of the curve (ridge) of the riches (economy) of China.
Mother tongue translation	Ukwehliswa kwentengo yemali yaseChina.	The decline of the price of money of China.	Thola amaphepha achaza izizathu, nezimbangela zokwehliswa kwentengo yemali yaseChina	Find papers that explain the reasons and causes for the decline of the price of money of China

Table A9.2 CLEF Query C052 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
China	1	hina	cut end shorten	There is no match for the proper name, since it only takes a class prefix.
	2	qhina	plait hair cramp movement tying obstruct put difficulties way	
	3	-qhina iqhina amaqhina	steenbok	
echaza	1	**chaza**	**make incisions tattoo incise order rub medicine reveal explain show meaning dominate crackle**	This is a match that literally means "to reveal the meaning".
	2	-chaza imichaza	greens	
	3	chachaza	dominate lord assume unrightful authority rude act respect drip fall drops crackle	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	A match was made, but note the easily mistranslated meanings like "large blanket".

	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
izizathu	1	**-zathu isizathu izizathu**	**reason excuse cause**	
	2	-zathu isizathu	sharp tongue	
	3	-thathu isithathu izithathu	third place third	
kwezinga	1	zinga	roam wander	This is an interesting form of a metaphor, where the meaning literally indicated a ridge (like the line of a graph) that declines.
	2	**-zinga izinga amazinga**	**ridge corrugation running elevation two closely placed grooves ridge two valleys annular ring tree scale snake fish tortoise millipede**	
	3	-zinga inzinga izinzinga	baboon roams	
lomnotho	1	**-notho umnotho**	**wealth riches soft springy object**	Homonym
	2	-nothonotho inothonotho izinothonotho	soft springy comfortable	
	3	-nothonotho isinothonotho izinothonotho	animal thick soft wool	
nemiphumela	1	-phumela umphumela imiphumela	small shrub love charm effect	
	2	phumela	come leave rise discharge relieve nature	
	3	-nomphumela unomphumela onomphumela	species bushveld shrub bearing fruit	
thola	1	**thola**	**find obtain pick come adopt receive family**	Homonym

	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukwehla	1	**-ehla ukwehla**	**slope declivity**	Note the explanation at kwezinga, where this words are indeed matched, but indicated a metaphor where a graph is illustrated.
	2	khehla	adopt head ring adopt top knot fell tree sit idle	
	3	-khehla ikhehla amakhehla	man head ring man beginning grey	
yokwehla	1	**-ehla ukwehla**	**slope declivity**	
	2	khehla	adopt head ring adopt top knot fell tree sit idle	
	3	-khehla ikhehla amakhehla	man head ring man beginning grey	

Table A10.1 CLEF Query C055 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Swiss Initiative for the Alps***		Find documents that report on the Swiss initiative aimed at regulating traffic through the Alps	
Official translation	Amasu eSwiss mayelana ne-Alps.	Plan of the Swiss with regards to the Alps.	Thola imibhalo echaza ngamasu eSwiss okuhlela ezemigwaqo e-Alps.	Find documents that report on the plan of the Swiss that is best for the road in the Alps.
Mother tongue translation	Ubukhanda beSwiss nge-Alipi.	The initiative of the Swiss for the Alps.	Thola imibhalo ebika ngegalelo leSwiss ngenhloso yokulungisela izinqola umqwaqo e-Alipi.	Find documents that report on the initiative of the Swiss to repair/put in order the wagons in the Alps

Table A10.2 CLEF Query C055 Matching

Word in official translation		Dictionary entry	Translations	Comment
Alps	1	lesi	this	There is no match for a proper name, it only takes a class prefix.
	2	-lisa isilisa	male kind men folk semen	
	3	-lisa umlisa abalisa	male person able daring man	
Swiss	1	-wisa iwisa amawisa	large headed knobkierie	There is no match for a proper name, it only takes a class prefix.
	2	wisa	cause fall fell knock mislead lead astray corrupt	
	3	-wisa isiwisa iziwisa	good snuff good beer hail medical charm keep hail gardens	
amasu	1	**-su isu amasu**	**plan method scheme ring cows horn succession birth**	Another homonym. The match is made to "plan", but could easily have translated to "a cow's horn" or "birth".
	2	-masu	relative word found compounds	
	3	-suku usuku izinsuku amasuku	day period twenty four hours period daylight	
ebika	1	**bika -vika**	**report introduce warn omen**	Homonym
	2	-bika ibika amabika	harbinger omen term applied bird insect species ant white spots introduction	
	3	-bibika imbibika izimbibika	cry baby mournful faced person	
ezemigwaqo	1	**-gwaqo umgwaqo imigwaqo**	**main large beaten track highway important places road wagon road route**	
	2	-gwaqo ingwaqo izingwaqo	difficulty	

	3	-gwaqa igwaqa amagwaqa		
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibalo**	**document writing mark entry scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
mayelana	1	**mayelana**	**direction opposite regard respect**	
	2	mayela	vicinity	
	3	-mayela imayela amamayela	mile	
ngamasu	1	**-su isu amasu**	**plan method scheme ring cows horn succession birth**	Homonym
	2	angama	lean overhang overlook cause reverance respect	
	3	ongama	lean overhang overlook cause reverance respect	
okuhlela	1	**okuhle**	**good luck best**	The word is formed through the verbal extension -ela added.
	2	fuhlela	fling leave disorder	
	3	-fuhlela imfuhlela izimfuhlela	collection rubbish untidy mass surging crowd	
thola	1	**thola**	**find obtain pick come adopt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	

Table A11.1 CLEF Query C056 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***European Campaigns against Racism***		Find documents that talk about campaigns against racism in Europe	
Official translation	Imikhankaso ye-Europe elwa nobandlululo.	The horseshoe-formation of Europe that struggles with discrimination.	Thola imibhalo ekhuluma ngemikhankaso yokulwa nobandlululo kwelase-Europe.	Find documents that talk about the "horseshoe-formation" that is struggling with discrimination in Europe.
Mother tongue translation	Isu lokwenza amalungiselelo ngokumelene nobandlululo ngezobuhlanga eYurobhu.	The making of arrangements that defend discrimination according to races in Europe.	Thola imibiko ekhuluma ngamalungiselelo ngokumelene nobandlululo eYurobhu.	Find reports that talk about the arrangements that defend discrimination in Europe.

Table A11.2 CLEF Query C056 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Europe	1	-kopela ukopela okopela	mean underhanded person sneak	There is no match in this case, since the proper name only take the class prefix.
	2	-popo upopo opopo ipopo	pawpaw	
	3	-peni upeni openi	threepenny piece	
ekhuluma	1	**khuluma**	**speak talk talk sense speak weight**	Homonym – the word could also have translated to "weight".
	2	-fezi- umfezi- abafezi	person ready tongue	
	3	-mamekhulu umamekhulu omamekhulu	grandmother	

elwa	1	-celwa ucelwa izincelwa	calabash	No match was made, since the word is a passive formed from a causitive verb stem by adding a -w-.
	2	felwa	lose death bereaved	
	3	-felwa umfelwa abafelwa	widower	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibal**	**writing document entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
imikhankaso	1	**-khankaso umkhankaso imikhankaso**	**horseshoe formation**	This is a very interesting and important match – the word campaign is matched metaphorically to "horseshoe formation" which is exactly what a campaign in Zulu terms is – normally in the context of a fighting campaign.
	2	-khankanyo umkhankanyo imikhankanyo	thought idea intention	
	3	-khankatho umkhankatho imikhankatho	scolding	
ngemikhankaso	1	**-khankaso umkhankaso imikhankaso**	**horseshoe formation**	Metaphor
	2	-khankanyo umkhankanyo imikhankanyo	thought idea intention	
	3	-khankatho umkhankatho imikhankatho	scolding	

nobandlululo	1	-bandlululo isibandlululo izibandlululo	suspension membership	
	2	**-bandlululo ubandlululo**	**discrimination analysis**	
	3	-bandlulo ubandlulo	differentiation	
thola	1	**thola**	**find obtain pick come adopt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
yokulwa	1	-sulwa insulwa izinsulwa	species astringent bulbous plant lightning medicine	No match was made, since the word is a passive formed from a causitive verb stem by adding a -w-.
	2	-sulwa umsulwa imisulwa	clean neat person chaste moral person	
	3	-sulwa ubusulwa	chastity good morals innocence	

Table A12.1 CLEF Query C057 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Tainted-Blood Trial***		Find all information about the tainted blood trials in France including the sentences given by the court and the names of the people found guilty	
Official translation	Ukuthethwa kwecala legazi elonakele	To give judgement over what is wrong of the blood that is spoilt/damaged.	Thola yonke imibiko mayelana nokuthethwa kwecala legazi elonakele eFulansi kanye nezigwebo ezanikezwa inkantolo namagama abantu abafunywana benecala.	Find that is many the reports regarding the judgement over the wrong of the blood that is spoilt/damaged in France that includes the judgement that was passed by the court and the names of people that were found guilty
Mother tongue translation	Ukuhlolwa kwegazi elonakele	To perform the test of the blood that is spoilt/damaged.	Thola lonke ulwazi ngokuhlolwa kwegazi elonakele eFrance kanye nesigwebo esakhishwa yinkantolo nabantu abatholwa banecala.	Find that is many the knowledge/thorough understanding of the test performed of the blood that is spoilt/damaged in France that includes the judgement that was passed by the court of people that was found to be wrong.

Table A12.2 CLEF Query C057 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
France	1	-ncengancenga incengancenga izincengancenga	fretful peevish child person needs humouring	This is a proper name that does not match to any entry in the dictionary, (although there is an entry for the Zululised eFulansi), and thus only take the appropriate class prefix.
	2	nce	tinkling metallic sound ticking	
	3	nce	order	

abafunywana	1	-dunywana udunywana odunywana	pin	There is no match since the word is formed from aba+bafana+ana. Palatalisation occurs (n<nyw), as well as vowel elision aba+bafana. Furthermore a verbal extension (-ana) is added to the stem.
	2	-dunywana umdunywana imidunywana	species forest myrtle tree	
	3	-fungwa umfungwa abafungwa	called oath	
abantu	1	**-ntu umuntu abantu**	**man person human member African Native race black man servant subject dependant person human feelings**	
	2	bantu	people say wonderful dear believe	
	3	-nta- umnta- abanta	child	
benecala	1	ngcala	play	No match wsa found, since the word (aba+ana+cala) was formed through vowel coalescence (a+i=e).
	2	enela	satisfy please keep force	
	3	-ncala incala	dirt	
elonakele	1	-nqakelela unqakelela onqakelela	eavesdropper	No match, since the word is formed through vowel coalescence (a+i=e), and the verbal extension -ele that was added.
	2	nqakelela	spread scandal divulge secrets eavesdrop	
	3	fakelela	replace fill make deficiency add replant fill spaces crops	
ezanikezwa	1	**nikeza**	**hand pass message recommend**	
	2	hlinikeza	rub put blame	
	3	shinikeza	rub roughly scrape scrape graze roughly	
imibiko	1	**-biko umbiko imibiko**	**report message announcement**	
	2	-boko umboko imiboko	elephants trunk penis horse proboscis insect	
	3	-bino umbino imibino	obscene language	

inkantolo	1	**-kantolo inkantolo izinkantolo**	**magistrates court charge office**	
	2	-santolo isisantolo izisantolo sántu	species shrubby climber tough thing easily torn broken	
	3	-ntolo intolo intol'ebomvu	weaknes knees species shrub	
kanye	1	kanye	way	No match was made, since the word is an enumerative formed by a+nye.
	2	kanyekanye	time	
	3	-nkanyezi inkanyezi izinkanyezi	star firefly	
kwecala	1	**-cala icala amacala**	**wrong deserving complaint defect mistake error fault crime offence law guilt guiltiness blame responsibility wrong charge law case law suit debt**	Homonym
	2	khweca	bend draw cause curve finish last remains	
	3	-zekecela izekecela amazekecela	scaly knobs row knobs crocodile skin	
legazi	1	**-gazi igazi amagazi**	**blood**	
	2	-gazi ingazi izingazi	blood shed clotted vitals	
	3	-gazi ugazi	personality commanding respect handsomeness attractiveness fine appearance	
mayelana	1	**mayelana**	**direction opposite regard respect**	
	2	mayela	vicinity	
	3	-mayela imayela amamayela	mile	
namagama	1	**-gama igama amagama**	**name appellation word letter alphabet statement song hymn air words song notability notoriety**	Although a match was made, the homonym could easily have been mistranslated.

	2	-gaga igaga amagaga	thorax mere skeleton bones emaciated person active spirited performance dash	
	3	-gamu igamu amagamu	lump flesh ribs pelvis space caused joint bending	
nezigwebo	1	**-gwebo isigwebo izigwebo**	**decision judgement hooked instrument**	Homonym
	2	-gwebo umgwebo imigwebo	upward rip sword thrust ripping stab	
	3	-gweba isigweba izigweba	bad blood heat supposed body Native child inordinate lust male female	
thola	1	**thola**	**find obtain pick come adopt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuthethwa	1	-thethe ubuthethe	mites maggots found old sour milk scurf dandruff	No match, since the stem -thetha changed to -thethwa through palatalisation.
	2	-phethwa iphethwa eng	game touch run	
	3	-khethe ukhethe izinkethe	laminated stone shale slates galvanized iron corrugated iron zinc	
yonke	1	**yonke**	**whole**	No exact match was made, since the word is a inclusive quantifier from Class 4 that means "that is many".
	2	onke	whole	
	3	konke	An inclusive quantifier in Class 15.	

Table A13.1 CLEF Query C058 Translations

	Query	Translation	Description	Translation
Original English CLEF query	Euthanasia		Documents will describe incidents of euthanasia understood as "death with dignity" or "the right to die"	
Official translation	Ukufa ngaphandle kokuzwa ubuhlungu.	*Death that is right of the senses that are painful.*	Imibhalo iyochaza izehlakalo zokufa ngaphandle kokuzwa ubuhlungu ezaziwa njengokufa okunesizotha noma ilungelo lokufa	Documents will explain the events of death that is right of the senses that is painful that is known as "to die with dignity" or "the right to die".
Mother tongue translation	Ukufa ngaphandle kokuzwa ubuhlungu	Death that is right of the senses that are painful.	Imibhalo izochaza izehlakalo zokufa ngaphandle kokuzwa ubuhlungu ezizwisiswa njengokufa ngesithunzi noma ilungelo lokufa.	Documents will explain the events of death that is right of the senses that is painful that is known as "to die without a shadow"/"to die with influence" or "the right to die"

Table A13.2 CLEF Query C058 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
ezaziwa	1	**aziwa**	**known understood**	
	2	-zazisa	proud opinion oneself	
	3	-azi isazi izazi	knows thing thoroughly expert scientist	
ilungelo	1	-lungelo ilungelo amalungelo	right privilege	
	2	-lungelunge ulungelunge izilungelunge	long parallel lines objects	
	3	-lunge ulunge izilunge	long row objects	

imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
iyochaza	1	**chaza**	**make incisions tattoo incise order rub medicine reveal explain show meaning dominate crackel**	Homonym that could also translate to "tattoo", "to rub medicine" or "to make an incision".
	2	-chaza imichaza	greens	
	3	chachaza	dominate lord assume unrightful authority rude act respect crackle	
izehlakalo	1	**-ehlakalo isehlakalo izehlakalo**	**event occurrence happening**	
	2	ehlakala	happen come pass occur	
	3	-hlakala isihlakala izihlakala	wrist	
kokuzwa	1	-khuzwa inkuzwa izinkuzwa	species aromatic tree medicinal tea perfume	Note the vowel coalescence (a+u=o) and vowel elision (uku+uzwa) in this match.
	2	-khuzwa umkhuzwa imikhuzwa	species aromatic tree	
	3	**-zwa uzwa izinzwa**	**sense feeling senses**	
lokufa	1	-loku		
	2	**-fa ukufa**	**death sickness disease**	
	3	ukufa	magnificent wonderful	
ngaphandle	1	**ngaphandle**	**externally right independently**	

	2	phandle		
	3	-phandle amaphandle	suburbs places habitation main village outskirts main kraal people live suburbs people acquired local customs	
ngokufa	1	ngokusa		No match, since the word is formed through vowel coalescence (a+u=o).
	2	ngokuba	dawn	
	3	ngoku-	form adverbs manner qualificatives	
nje	1	nje	manner	No match, since the word is known as an enclitic. It does not have any meaning.
	2	-nje	merely perfectly	
	3	-lenje ilenje amalenje	volunteer member rifle association	
noma	1	**noma**	**if although**	
	2	oma	dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
okunesizotha	1	**-zotho isizotho**	**strong personality dignity**	Matched to correct dictionary entry, although the word forms part of the parahrase "to die with dignity".
	2	zotha	calm mellow sober state appearance	
	3	-zotha inzotha izinzotha	beast marked sober colouring	
ubuhlungu	1	-hlungu ubuhlungu	pain herb	
	2	**-buhlungu**	**painful**	
	3	kabuhlungu	painfully pitifully touchingly	
ukufa	1	**-fa ukufa**	**death sickness disease**	
	2	ukufa	magnificent wonderful	
	3	fakaza ukufakaza	give evidence bear witness testify give testimony	

zokufa	1	**-fa ukufa**	**death sickness disease**
	2	ukufa	magnificent wonderful
	3	-zwekufa izwekufa amazwekufa	state desolation ruin

Table A14.1 CLEF Query C059 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	***Computer viruses***		Find documents about computer viruses	
Official translation	Amavayirasi ekhompiyutha.	Virus of the computer.	Thola imibhalo mayelana namavayirasi wekhompiyutha.	Find documents regarding viruses of the computer.
Mother tongue translation	Ubuthi bama-"computer".	Poison/virus of the computer.	Thola imibhalo ngobuthi bama-"computer".	Find documents about poison of the computer.

Table A14.2 CLEF Query C059 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
ama	1	ama-	noun prefix class three plural	No match in this case, since the match is incorrectly being made to a class prefix that actually forms part of a Zululised term for virus.
	2	ama-	adjective concord class three plural	
	3	-ama	verbial derivative suffix	
computer	1	-putumputu imputumputu izimputumputu	crumbly nature	No match would be made to this word, since it is a borrowed word that only takes a class prefix (ama- from class 6).
	2	-phumputhe impumputhe izimpumputhe	blind person	
	3	-kaputeni ukaputeni okaputeni	captain	

imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibalo**	**document writing mark entry scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
mayelana	1	**mayelana**	**direction opposite regard respect**	
	2	mayela	vicinity	
	3	-mayela imayela amamayela	mile	
nama	1	enama	happy contented blissful exilerated state pleased complacent glad	No match is possible, since the word is part of a class prefix that is added to virus (a borrowed word in a zululised form).
	2	-manamana amamanamana	quibbling subterfuge	
	3	ncama	eat commencing journey satisfy prefer lose hope give	
thola	1	**thola**	**find obtain come pick adobt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
viruses	1	-risese urisese orisese	recess interval school lessons break	No match is possible, since the word is a borrowed word that are also zululised.
	2	-zibuse uzibuse ozibuse	responsible government	
	3	-sese isese	secret place place hiding	

Table A15.1 CLEF Query C060 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Corruption in French politics***		Find documents on corruption in politics in France, in particular with reference to the illegal financing of French political parties	
Official translation	Ukonakala kwezepolitiki kwelaseFulansi.	That which is corrupt in the politics of France.	Thola imibhalo ngokonakala kwezepolitiki eFulansi ikakhulukazi maqondana nokuzuzwa ngokungemthetho kwemali yizinhlangano zepolitiki zaseFulansi	Find documents about that which is corrupt in the politics of France specially that which refers to the condition according to law of money in the agreement with politics of France.
Mother tongue translation	Inkohlakalo kwezepolitiki eFrance.	Corruption in the politics of France.	Thola imibhalo ngenkohlakalo kwezepolitiki eFrance, ikakhulukazi maqondana nokungavunyelwe kokuxhasa ngemali kwamaqembu epolitiki aseFrance.	Find documents about corruption of the politics of France, especially that which refers to that which is permitted to bring pressure on money of the divisions of politics in France.

Table A15.2 CLEF Query C060 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
France	1	-ncengancenga incengancenga izincengancenga	fretful peevish child person needs humouring	A proper name that only takes a class prefix.
	2	nce	tinkling metallic sound ticking	
	3	nce	order	
ikakhulu	1	**yikakhulu**	**especially principally**	
	2	kakhulu	greatly loudly chiefly certainly	

	3	-makhulu umakhulu omakhulu	grandmother	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibalo**	**document writing mark entry scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
izinhlangano	1	**-hlangano inhlangano izinhlangano**	**meeting assembly league agreement alliance pact**	
	2	-hlanga inhlanga izinhlanga	thicket reeds incision incised pattern brand trade mark	
	3	-hlanga uhlanga izinhlanga	dry stalk reed snuff box tube original stem stock ancestry genealogy tribal incision medical incision	
kwemali	1	**-mali imali izimali**	**money cash**	
	2	gwema	avoid stand aside way	
	3	-ali umali abali	refuses rejects dislikes	
kwezepolitiki	1	-wositiki uwositiki owositiki	walking stick	No match because the word is in a zululised form.
	2	-kweleti isikweleti izikweleti	debt	
	3	-dibikezelo isidibikezelo izidibikezelo	lid	
maqondana	1	**maqondana**	**line straight direction opposite reference appertaining**	
	2	qondana	understand opposite line	
	3	-qondo iqondo amaqondo	stitch single binding directions administering medicine stone bladder	

ngokonakala	1	**onakala**	**spoilt injured damaged corrupted depraved wrong miscarry**	
	2	bonakala	appear come vision visible obvious evident clear plain revealed found	
	3	ngonkana	whole	
ngokungemthetho	1	**ngomthetho**	**right lawfully according law constitutionally**	
	2	-thetho umthetho imithetho	law custom rule edict statute pronouncement arrangement	
	3	-thetho intetho izintetho	subject matter speech utterance	
nokuzuzwa	1	**zwa yizwa ukuzizwa**	**perceive various senses hear listen ear obey understand taste smell feel touch live alive sound good condition feel self important**	
	2	-zuzwana umzuzwana imizuzwana	moment short period time	
	3	-khuzwa inkuzwa izinkuzwa	species aromatic tree	
thola	1	**thola**	**find obtain come pick adobt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukonakala	1	**onakala**	**spoilt injured damaged corrupted depraved wrong miscarry**	
	2	bonakala	appear come vision visible obvious evident clear plain revealed found	
	3	-onakalo umonakalo	mishap mess upsetting plans	
zepolitiki	1	-wositiki uwositiki owositiki	walking stick	No match was correctly made because the word is a zululised form.

	2	-poli isipoli izipoli	railway line	
	3	-tiki utiki otiki	three penny piece tickey	

Table A16.1 CLEF Query C062 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Northern Japan earthquake***		Find reports that report on an earthquake on the east coast of Hokkaido, northern Japan, in 1994	
Official translation	Ukuzamazama komhlaba eNyakatho yeJapani.	The shaking of the earth of the North of Japan.	Thola imibhalo ebika ngokuzamazama komhlaba ogwini lasempumalanga yeHokkaido yeNyakatho yeJapani ngo-1994.	Find documents that report about the shaking of the earth at the edge of the water of the east of Hokkaido of the North of Japan in 1994.
Mother tongue translation	Ukuzamazama komhlaba eNyakatho yaseJapan.	The shaking of the earth of the North of Japan.	Thola imibhalo ebika ngesehlakalo sokuzamazama komhlaba ogwini lwasempumalanga yeHokkaido, eNyakatho yaseJapan ngo-1994.	Find documents that report about the demolishing of the shaking of the earth at the edge of the water of the east of Hokkaido of the North of Japan in 1994.

Table A16.2 CLEF Query C062 Matching

Word in official translation		Dictionary entry	Translations	Comment
Hokkaido	1	fokoka	bashed considerably dented springy pressed crackle pressure	No match is made because the word is a proper name.
	2	wokoka	disintegrate fall pieces decompose soft flabby	
	3	koka	encircle surround	
Japan	1	-kapane isikapane izikapane	evil smell evil smelling object	No match is made because the word is a proper name.

	2	pana	hobble knee halter treat strict discipline	
	3	-pani ipani amapani	frying pan saucepan	
ebika	1	**bika -vika**	**report introduce announce warn omen**	
	2	-bika ibika amabika	harbinger omen term applied insect bird species ant	
	3	-bikado umbikado imibikado	spear	
enyakatho	1	**enyakatho**	**north**	
	2	-nyakatho inyakatho	North North wind	
	3	-bekenyakatho ubekenyakatho obekenyakatho	looks north white iron wood	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
komhlaba	1	-hlaba umhlaba imihlaba	species aloe	This word must be read together with ukuzamazama in order to contextualise the meaning of them – earthquake – literally translated to the paraphrased metaphor "the shaking of the earth".
	2	**-hlaba umhlaba imihlaba**	**earth world land country value**	
	3	-hlababa umhlababa	species thorny shrub	
lasempumalanga	1	**empumalanga**	**east**	
	2	-phumalanga impumalanga	east sun comes	

	3	-galanga umgalanga imigalanga	species tree resembling waterboom	
ngo-1994	1	ngo-	excessive heat tight grip strong denial firmness	There is no match in this case, since the numeral is preceded by an inflected prefix.
	2	-ongo isongo izongo	sparing economy	
	3	-ongo isongo izongo	uncooked centre food momentary glare sun	
ngokuzamazama	1	**zamazama**	**shake vibrate quake rock fro**	This word must be read together with komhlaba in order to contextualise the meaning of them – earthquake – literally translated to the paraphrased metaphor "the shaking of the earth".
	2	ngokusa	dawn	
	3	ngokuba		
ogwini	1	-wini	locative suffix form locative adverbs nouns ending vowel u	There is no match for this locative form, since it forms part of "coast" that is translated literally to the paraphrased metaphor "at the edge of the water".
	2	-wini umwini abawini	winner Ricksha boy	
	3	-gwinja ugwinja ogwinja	diver	
thola	1	**thola**	**obtain find come pick adopt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuzamazama	1	**zamazama**	**shake vibrate quake rock fro**	This word must be read together with komhlaba in order to contextualise the meaning of them – earthquake – literally translated to the paraphrased metaphor "the shaking of the earth".
	2	nkazama	afternoon	
	3	-nkazama inkazama	afternoon	
yenyakatho	1	**enyakatho**	**north**	
	2	-nyakatho inyakatho	North North wind	

	3	-bekenyakatho ubekenyakatho obekenyakatho	looks north white iron wood	

Table A17.1 CLEF Query C063 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Whale reserve***		Find documents about the reserve in the Antarctic in which hunting for whales is forbidden	
Official translation	Isiqiwu semikhoma.	Boundary marked for whales.	Thola imibiko mayelana nesiqiwu e-Antatika lapho ukuzingelwa kwemikhoma kungavunyelwe khona.	Find documents regarding the boundary marked in Antarctica where the hunting of whales is not permitted
Mother tongue translation	Indawo eyabelwe izilwane ezinkulu zasolwandle.	The place that is reserved for animals that is big of the sea.	Thola imibhalo emayelana nedawo eyabiwe eAntarctic lapho kunguvunyelwe khona ukuzingelwa kwezilwane zasolwandle.	Find documents regarding the place that is reserved in the Antarctica where it is not permitted there to hunt animals of the sea.

Table A17.2 CLEF Query C063 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Antarctic	1	-thacataca intacataca izintacataca	soft adhesive mud	No match was possible, since the proper name only takes the class prefix to which it belongs.
	2	-ntaca intaca izintaca	snake	
	3	-fanta umfanta imifanta	crack	
imibiko	1	**-biko umbiko imibiko**	**report message announcement**	

	2	-boko umboko imiboko	elephants trunk penis horse proboscis insect	
	3	-bino umbino imibino	use obscene language	
isiqiwu	1	**-qiwu isiqiwu iziqiwu**	**beacon boundary mark**	An interesting match again from a cultural point of view, since the word is paraphrased from "a boundary marked for" that translates to the reserve.
	2	-qimu isiqimu	Rinderpest Bubonic plague	
	3	-nqiwu inqiwu izinqiwu	plan plot conspiracy	
kungevunyelwe	1	**vunyelwa**	**permitted**	
	2	-dunyelwa umdunyelwa abadunyelwa	person assaulted complainant assault case	
	3	-lungelunge ulungelunge izilungelunge	long parallel lines objects	
kwemikhoma	1	**-khoma umkhoma imikhoma**	**whale**	Interestingly enough, in the mother tongue translation the word whale is paraphrased to "animals of the sea that is big"
	2	-khomakhoma umkhomakhoma imikhomakhoma	tree fern	
	3	-khomazi umkhomazi imikhomazi	female whale whale cow	
lapho	1	lapho	there	
	2	**lapho**	**where**	
	3	lapho	time	
mayelana	1	**mayelana**	**direction opposite**	
	2	mayela	vicinity	
	3	-mayela imayela amamayela	mile	

nesiqiwu	1	**-qiwu isiqiwu iziqiwu**	**beacon boundary mark**	
	2	-nqiwu inqiwu izinqiwu	plan plot conspiracy	
	3	-nqiwu umnqiwu	grudge persistent ill feeling	
semikhoma	1	**-khoma umkhoma imikhoma**	**whale**	
	2	-khomakhoma umkhomakhoma imikhomakhoma	tree fern	
	3	-khomazi umkhomazi imikhomazi	female whale whale cow	
thola	1	**thola**	**find obtain come pick adobt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuzingelwa	1	**zingela**	**roam hunt search lie wait pursue**	In this correct match, the word is formed through pre-nasalisation where the -w- is added before the last vowel.
	2	zingeleza	encircle surround circulate	
	3	-zingeli umzingeli abazingeli	hunter huntsman pursuer	

Table A18.1 CLEF Query C064 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Computer mouse RSI***		Find documents that report on computer mouse repetitive strain injuries (RSI)	
Official translation	Ukusetshenziswa kwama-"computer mouse" ngokuphindaphindiwe okuletha ukulimala kwama-"computer mouse".	To cause by the computer mouse that which is repeated to bring injury of the mouse.	Thola imibiko ebika ngokuphindaphindeka kwezikhubazo zokuqanwa isandla ngokusetshenziswa kwegundana lekhompiyutha.	Find documents that report about the causes of that which is repeated the injuries of stress of the hands when using the mouse of the computer
Mother tongue translation	Igundane lekhompiyutha lokusetshenziswa ngokuphindaphindiwe okuletha kokulimala.	Mouse of the computer that causes that which is repeated to bring injuries.	Thola imibhalo ebika ngokhuphindaphindeka okungenelele kokulimala kwegundane lekhompiyutha.	Find documents that report on that which is repeated to bring injury of the mouse of the computer.

Table A18.2 CLEF Query C064 Matching

Word in official translation		Dictionary entry	Translations	Comment
computer	1	-putumputu imputumputu izimputumputu	dry crumbly nature	This is a borrowed word and would not match to any dictionary entry. Interestingly enough is the fact that the zululised form (-ikhompiyutha) appears in the mother tongue translation.
	2	-phumputhe impumputhe izimpumputhe	blind person	
	3	-kaputeni ukaputeni okaputeni	captain	
ebika	1	**bika -vika**	**report introduce warn omen**	
	2	-bika ibika amabika	harbinger omen term applied insect bird	
	3	-bibika imbibika izimbibika	cry baby mournful faced person	

imibiko	1	**-biko umbiko imibiko**	**report message announcement**	
	2	-boko umboko imiboko	elephants trunk penis horse proboscis insect	
	3	-bino umbino imibino	obscene language	
kwama	1	-khwama isikhwama izikhwama	pouch made cows bladder small bag purse pocket pouch difficulty	This word would not match any dictionary entry, since it is the class prefix that is added to the borrowed word computer.
	2	kama	comb	
	3	kwamhlá kwami	day	
mouse	1	-mese umese omese	knife	This is a borrowed word to which a class prefix is added, thus no correct match would be found. Interestingly from a cultural point of view the mother tongue directly translates the word to –igundane that means "mouse". It is however not in the context of the computer world.
	2	kuse	act	
	3	-phusela impusela izimpusela	calf ceased sucking weaned calf	
ngokuphindaphindiwe	1	-phindlaphindla uphindlaphindla izimpindlampindla	person animal dodges tall lanky person	In this case the duplicated verb form is matched to the single verb entry, with a verbal extension occurring in the end.
	2	-daphundaphu indaphundaphu izindaphundaphu	soft natured ungraspable thing	
	3	**phinda**	**repeat again return back fold make double**	
ngokusetshenziswa	1	ngokushesha	quickly hurriedly	Note the palatalisation that occurs in this example.
	2	**enziswa**	**cause help make**	
	3	tshengisa	show demonstrate point	
okuletha	1	**leso letha**	**bring**	

	2	-letha iletha amaletha	species small tree	
	3	-lethi ulethi olethi	species shrub growing near rivers	
thola	1	**thola**	**find obtain come pick adobt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukulimala	1	**limala**	**hurt damaged suffer receive injury disappointed inconvenienced**	This word is actually paraphrased to "to receive injury".
	2	-lima ubulima	deformity malformation idiocy foolishness stupidity	
	3	-lala umlala imilala	vertebral ligament	
ukusetshenziswa	1	**enziswa**	**cause help make**	Although a match is made to the verb stem, palatalisation occurs through b>tsh.
	2	- ketshe uketshe izinketshe	thin watery substanc	
	3	tshengisa	show demonstrate point	

Table A19.1 CLEF Query C065 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Treasure hunting***		Find documents about treasure hunters and treasure hunting activities	
Official translation	Ukuzingelwa kwamagugu eminotho.	To hunt that which is precious/valuable that is of riches.	Thola imibhalo ngabazingeli bamagugu omnotho kanye nezenzo zokuzingelwa kwamagugu omnotho.	Find documents about hunters of valuables of wealth and the way and actions to hunt valuables of wealth.
Mother tongue translation	Ukuzingelwa kwezinto ezingamagugu.	To hunt for things that is of value/precious.	Thola imibhalo mayelana nokuzingelwa kwezinto eziyigugu kanye nobukhuphekhuphe bokufuna izinto eziyigugu.	Find documents regarding the hunting of things that is precious and the way and activities to look for things that are precious.

Table A19.2 CLEF Query C065 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
bamagugu	1	**-gugu igugu amagugu**	**valuable prized rare object favourite person precious stone**	The match is made to the stem, which actually forms part of the paraphrased "that which is precious/valuable". This is a literal translation of treasure.
	2	-gugo igugo amagugo	old man	
	3	-gagu igagu amagagu	bold forward person big talker boastful person expert music speaking	
eminotho	1	-nothonotho inothonotho izinothonotho	soft springy comfortable	This word actually forms part of the paraphrase for treasure, and must be read in context with "that which is precious/valuable".
	2	**-notho umnotho**	**wealth riches soft springy object**	
	3	-nothonotho isinothonotho izinothonotho	animal thick soft wool	

imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
kanye	1	kanye	way	There is no match, since the word is an enumerative formed by a+nye.
	2	kanyekanye	time	
	3	-nkanyezi inkanyezi izinkanyezi	star firefly	
kwamagugu	1	**-gugu igugu amagugu**	**valuable prized rare object favourite person precious stone**	
	2	-gugwa igugwa amagugwa	species herb	
	3	-gagu igagu amagagu	bold forward person big talker boastful person expert music speaking	
nezenzo	1	-nenezo umnenezo iminenezo	short skin petticoat worn women menstruation	The incorrect matches in this case is because of changes to the stem (-izenzo) through vowel coalescence (a+i=e).
	2	-zenze izenze amazenze	flea	
	3	-zenze isizenze izizenze	scissors shears battle axe	
ngabazingeli	1	**-zingeli umzingeli abazingeli**	**hunter huntsment pursuer**	
	2	-gaba ingaba izingaba	large branch bough	
	3	zingela	roam roam hunt search lie wait pursue	
omnotho	1	**-notho umnotho**	**wealth riches soft springy object**	Homonym

	2	-nothonotho inothonotho izinothonotho	soft springy comfortable	
	3	-nothonotho isinothonotho izinothonotho	animal thick soft wool	
thola	1	**thola**	**find obtain come pick adobt receive family**	Homonym
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuzingelwa	1	**zingela**	**roam hunt search lie wait pursue**	Some palatalisation occurs through the -w- being added before the last vowel.
	2	zingeleza	encircle surround circulate	
	3	-zingeli umzingeli abazingeli	hunter huntsman pursuer	
zokuzingelwa	1	**zingela**	**roam hunt search lie wait pursue**	
	2	zingeleza	encircle surround circulate	
	3	-zingeli umzingeli abazingeli	hunter huntsman pursuer	

Table A20.1 CLEF Query C067 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Ship collisions***		Find information on the number of people injured or killed in collisions between ships	
Official translation	Ukushayisana kwemikhumbi.	To hit each other that is ships.	Thola umkhondo ngesibalo sabantu abalimele noma abafe engozini yokungqubuzana kwemikhumbi	Find reports about the number of people that are hurt/received injury or died in that was to hit each other in the middle of ships.
Mother tongue translation	Ukungqubuzana kwemikhumbi.	The collision of ships.	Thola umkhondo ngesibalo sabantu abakhubazeka nabafayo ngokushayisana kwemikhumbi.	Find knowledge/a thorough understanding about the number of people that were hurt/injured and killed that was to hit each other in the middle of ships.

Table A20.2 CLEF Query C067 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
abalimala	1	**limala**	**hurt damaged suffer receive injury**	
	2	-lima ilima amalima	communal cultivation working bee cultivating field ploughed communal cultivation	
	3	-limi umlimi abalimi	farmer gardener	
abashona	1	**-sho umusho abasho**	**person**	
	2	shobashoba	wriggle wriggle move restlessly fidget	
	3	-shobashoba isishobashoba izishobashoba	restless person	

imibiko	1	**-biko umbiko imibiko**	**report message announcement**	
	2	-boko umboko imiboko	elephants trunk penis horse proboscis insect	
	3	-bino umbino imibino	obscene language	
kwemikhumbi	1	**-khumbi umkhumbi imikhumbi**	**assembly men drawn circle long narrow trough thing ship steamer rope**	A very good example of a homonym, where the word could easily have been mistranslated to "long narrow trough", or "an assembly of men" or "rope".
	2	-khumbi ikhumbi	sage grass thatching	
	3	-khumbi isikhumbi izikhumbi	nation	
ngesibalo	1	**-balo isibalo izibalo**	**figure cipher number mathematical problem**	
	2	-balo ibalo amabalo	numerical order status	
	3	ngesihle	freely gratis	
ngokushayisana	1	**shayisa**	**cause strike help hit knock work finish work day**	A correct match is made, although the verb changes through the verbal extension (-ana) being added.
	2	ngokusa	dawn	
	3	ngokushesha	quickly hurriedly	
noma	1	noma		No match for this conjunction (na+uma) meaning "whether...or".
	2	oma	dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
phakathi	1	**phakathi**	**midst middle**	In a cultural context, it is important to note that the translation actually attempts to describe exactly where the ships hit each other.
	2	-phakathi iphakathi amaphakathi	centre	

	3	-phakathi umphakathi imiphakathi	assembly men district commoners apex arch female genitals	
sabantu	1	**-ntu umuntu abantu**	**human being person man member African Native race black man person human feelings subject servant dependant**	
	2	bantu	people say wonderful dear believe	
	3	-nta- umnta- abanta	child	
thola	1	**thola**	**obtain find pick come adopt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukushayisana	1	**shayisa**	**cause strike help hit knock work finish work day**	A correct match is made, although the verb changes through the verbal extension (-ana) being added.
	2	shayela ukushayela	strike punish hammer finish timbering drive inspanned harnessed animals drive whip drive motor car	
	3	shayela ukushayela	sweep brush away dance informally	

Table A21.1 CLEF Query C068 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Attacks on European Synagogues***		Find documents that describe acts of terrorism or vandalism against European synagogues since the end of the Second World War	
Official translation	Ukuhlaselwa kwezindlu zesonto zamaJuda aseYurobhu.	The attacks on the dwelling place of Sundays of the Jews of Europe.	Thola imibhalo echaza izenzo zobuphekulazikhuni noma zokwephula izinto ngobudlova, elwa nezindlu zesonto zamaJuda aseYurobhu, kusukela ekupheleni kwempi yesibili yomhlaba.	Find documents that explain the actions of militants/communists and the breaking of things violently against the dwelling place of Sundays of the Jews of Europe, that was undertaken at the end of the war that is two of the world.
Mother tongue translation	Ukuhlaselwa kwamasinagoge ase-Europa.	The attacks on the synagogues of Europe.	Thola imibhalo echaza kabanzi ngezehlakalo zokuthusela kanye nokoniwa kwezinto emasinagogeni ase-Europa kusukela ekupheleni kwempi yesibili zomhlaba.	Find documents that explain that of fear and the breaking down and the way to act as vandals of things of synagogues of Europe that was undertaken at the end of the war that is two of the world.

Table A21.2 CLEF Query C068 Matching

Word in official translation		Dictionary entry	Translations	Comment
Europe	1	-kopela ukopela okopela	mean underhanded person sneak	A proper name that would not match to any word in the dictionary. It only takes a class prefix.
	2	-popo upopo opopo ipopo	pawpaw	
	3	-peni upeni openi	threepenny piece	
Juda	1	-Juda iJuda amaJuda	Jew	Another proper name, although the Zululised form is matched to a dictionary entry.
	2	guda	milk cow giving calf suck milk dry milk drop question craftily order extract information clean interior scrape	

	3	duda	flatter indulge humour smile make conceited inspiration	
echaza	1	**chaza**	**make incisions tattoo incise order rub medicine reveal explain show meaning dominate crackle**	Homonym that could also mean "to make incisions", "tattoo" or "to rub medicine".
	2	-chaza imichaza	greens	
	3	chachaza	dominate lord others assume unrightful authority	
ekupheleni	1	**ekupheleni**	**end**	
	2	-sheleni usheleni osheleni	shilling	
	3	kuphela	merely alone	
elwa	1	-celwa ucelwa izincelwa	calabash	No match was made, since the word is a passive formed from a causitive verb stem by adding a -w-.
	2	felwa	lose death bereaved	
	3	-felwa umfelwa abafelwa	widower	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	Homonym
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
izenzo	1	**-enzo isenzo izenzo**	**deed action act verb**	
	2	-enzo ulwenzo izenzo		
	3	-zenze izenze amazenze	flea	
izinto	1	**-tho into izinto**	**thing object article matter affair property**	
	2	-tho utho izinto		
	3	-ntoyi intoyi izintoyi	red ivory tree	
kusukela	1	**sukela**	**away set originate grow tall spring jump attack undertake unprepared act due thought**	

	2	esukela	set originate grow tall spring jump attack undertake unprepared act due thought	
	3	-sekela usekela	distemper dog	
kwezimpi	1	**-mpi impi izimpi**	**regiment army military force encounter fight engagement battle war hostile person foe enemy**	
	2	-khwezi inkwezi izinkwezi	small star species shrub	
	3	-phici uphici izimpici	squashing crushing perspiring oozing fat slipping sliding grazing skin talking evasively	
kwezindlu	1	**-ndlu indlu izindlu**	**dwelling place habitation family descendants tribe**	This is a good example of paraphrasing, where the word must be read in context with "the dwelling place of Sundays of the Jews" – literally meaning a synagogue. (There is a word isinagoge in the dictionary that is a zululisation).
	2	-dlubu indlubu izindlubu	species underground nut	
	3	-dlubu udlubu izindlubu	single grain underground nut	
nezindlu	1	**-ndlu indlu izindlu**	**dwelling place habitation family descendants tribe**	As in the above description. It can also be added that this word is a homonym that could translate to "family descendants" or "tribe".
	2	-dluze indluze izindluze	flower Bottlebrush tree	
	3	-dlundlu indlundlu izindlundlu	self conceit haughty pride	
ngobudlova	1	-dlova ubudlova	ruffianism high handedness	In this example a match is made to the correct stem. However, the word forms part of a paraphrase that contextualise "the breaking of things violently".
	2	-gobandlovu ugobandlovu	species bushveld tree species forest climber tree stone gravel disease	
	3	**dlova**	**treat roughly violently override**	
noma	1	noma		No match for this conjunction (na+uma) meaning "whether...or".
	2	oma	dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
thola	1	**thola**	**obtain find come pick adopt receive family**	

	2	-thola ithola amathola eng	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuhlaselwa	1	**hlasela**	**war invade attack hunt**	
	2	hlwa ukuhlwa	dark dusk eclipsed mystified stupefied loss	
	3	-hlaseli umhlaseli abahlaseli	attacker invader hunter	
zesibili	1	-lwesibili ulwesibili	Tuesday	The word is match to the correct stem. However, the word forms part of the paraphrase for World War II, in the context of "the war that is two of the world".
	2	-bili isibili izibili	substance body real self personality truth reality essential parts body life	
	3	**-bili isibili**	**second place second**	
zesonto	1	**-sonto isonto amaso**	**Sunday week church building religious service denomination**	The match forms part of the paraphrase that means "dwelling place of Sundays of the Jews" that literally means synagogue.
	2	-sonto insonto izinsont	rope twisted calf skin Berlin wool woolen thread	
	3	-sonto umsonto imisont	thread twisted woven work	
zobuphekulazikhuni	1	**-phekulazikhuni iphekulazikhuni amaphekulazikhuni**	**Bolshevist militant communist**	
	2	phekula	blow upset overturn capsize cause change mind lure disaster mislead	
	3	-phekula isiphekula iziphekula	cheeky quarrelsome person	
zokwephula	1	cwephula	cut pieces	This word is formed through vowel coalescence (za+uku=zoku) and pre-nasalisation (uku+e=ukwe).
	2	khephula	break piece cut piece	
	3	**ephula**	**break cause sudden death disappoint**	

zomhlaba	1	-hlaba umhlaba imihlaba	species large aloe	The match must be read in the context of the paraphrase "war that is second of the world", that translates to World War II.
	2	**-hlaba umhlaba imihlaba**	**earth world land country value**	
	3	hlaba	pierce stab stick gore thrust kill slaughter attack pain reach perfectly hurt feelings wound mentally criticize review drive	

Table A22.1 CLEF Query C070 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	***Death of Kim II Sung***		Find documents giving biographical information on Kim II Sung, the president of North Korea, who died in 1994	
Official translation	Ukufa kukaKim II Sung	*Death of Kim II Sung*	Thola imibhalo enikeza ulwazi lomlando wempilo kaKim II Sung, umongameli wenyakatho eKorea, owashona ngo-1994.	Find documents that pass knowledge of the history of the life of Kim II Sung, the king of the North of Korea, that died in 1994.
Mother tongue translation	Ukukhothama kukaKim Sung wesibili [the word is used out of respect for when kings die]	To humbly bow down for Kim Sung that is second.	Thola imibhalo enikeza iminingwane ngoKim Sung wesibili owayengumongameli waseNyakatho neKorea, owashona ngo1994	Find documents that pass details of Kim Sung that is second, that looked after (be president) of the North of Korea that passed away in 1994

Table A22.2 CLEF Query C070 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
II	1	-thongo ithongo amathongo	ancestral spirit	No match was made because the word is a proper name that only takes a class prefix.
	2	babela	burn grass	

	3	babela	pungent bitter acrid burning desire consistently aim	
Kim	1	imá		No match was made because the word is a proper name that only takes a class prefix.
	2	im- in-		
	3	-dima idima amadima	cultivated field	
Korea	1	-roza iroza amaroza	rose rose bush	No match was made because the word is a proper name that only takes a class prefix.
	2	-hora ihora amahora	hour	
	3	reka	rake play rag time music engage movement African music	
Sung	1	-ungu umungu	fine dust flying corn sifted pollen	No match was made because the word is a proper name that only takes a class prefix.
	2	-gungu ingungu izingungu	enclosed shelter friction drum curved convex object	
	3	-gungu isigungu izigungu	secret plot scheme done secretly	
enikeza	1	**nikeza**	**hand pass message recommend**	
	2	hlinikeza	rub put blame	
	3	shinikeza	rub roughly scrape graze roughly	
eyashona	1	**shona**	**sink down sight disappear sunset die lose heavily poor bankrupt ruined**	This homonym could easily have been mistranslated to any of the other meanings.
	2	-nyashoba inyashoba izinyashoba	crupper	
	3	tshona	see	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	This homonym could easily have been mistranslated to any of the other meanings.
	2	**-balo umbalo imibalo**	**document writing entry mark scriptures large blanket**	

	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
inkosi	1	**-khosi inkosi amakhosi**	**king paramount chief term respect royalty magistrate spirits departed Lord**	This homonym could easily have been mistranslated to any of the other meanings.
	2	-nkosa inkosa izinkosa	filament styles female flower eng twt	
	3	-khosikazi inkosikazi amakhosikazi	principal wife chief headman lady	
lomlando	1	**-lando umlando imilando**	**narrative relation affair history**	In the cultural context, this word is metaphorically used to paraphrase "biographical" to "the history of the life"
	2	-lando isilando izilando	narrative	
	3	-landu umlandu imilandu	grudge blot character moral lapse history	
ngo-1994	1	ngo-	excessive heat tight grip strong denial firmness	There is no match in this case, since the numeral is preceded by an inflected prefix.
	2	-ongo isongo izongo	sparing economy	
	3	-ongo isongo izongo	uncooked centre food momentary glare sun	
thola	1	**thola**	**obtain find come pick adopt receive family**	This homonym could easily have been mistranslated to any of the other meanings.
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike meagre picking scarcity scantiness	
ukufa	1	**-fa ukufa**	**death sickness disease**	
	2	ukufa	magnificent wonderful	
	3	fakaza ukufakaza	indistinct markings drizzle weave embroider colour	

ulwazi	1	**-azi ulwazi**	**thorough knowledge understanding**	
	2	-klwazi inklwazi	rippling babbling sound sound womens voices singing	
	3	-hlwazi umhlwazi imihlwazi	species rare tree protea species green striped snake	
wempilo	1	-philo impilo	**health life mode living mode subsistence livelihood**	
	2	-swempe uswempe izinswempe	old worn out bull recluse	
	3	-swempe umswempe umiswempe	juicy semi transparent root plant	
yenyakatho	1	**enyakatho**	**north**	This word attempts to match North Korea, but only partially match the proper name North.
	2	-nyakatho inyakatho	North North wind	
	3	-bekenyakatho ubekenyakatho obekenyakatho	white iron	

Table A23.1 CLEF Query C071 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Vegetables, fruit and cancer***		Find documents that relate to the eating of vegetables and fruit to cancer	
Official translation	Imifino, izithelo nekhensa (dictionary translation for cancer = isigaxa or umhlaza)	*Vegetables, fruit and cancer*	Thola imibhalo elandisa ngokuxhumana kokudla imifino nezithelo kwikhensa.	Find documents that give an account of the eating of vegetables and fruit to cancer.
Mother tongue translation	Imifino, izithelo nomdlavuza	Vegetables, fruit and "that which tears apart" (cancer)	Thola imininingwane ephathelene nokudliwa kwemifino nezithelo kumdlavuza	Find documents that deal with the eating of vegetables and fruit to cancer.

Table A23.2 CLEF Query C071 Matching

Word in official translation		Dictionary entry	Translations	Comment
elandisa	1	**landisa**	**cause fetch imitate relating tell relate narrate give account**	
	2	andisa	increase cause spread cause flourish propagate enlarge	
	3	-landiso isilandiso izilandiso	predicative	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	
	2	**-balo umbalo imib**	**document writing entry mark scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
imifino	1	**-fino umfino imfino imifino**	**ordinary food green vegetables edible herb cooked food**	
	2	-fino isifino izifino	daily ordinary food kind tasty dish European food	
	3	-fingqo umfingqo imifingqo	down jerky movement	
izithelo	1	**-thelo isithelo izithelo**	**fruit crop deed product work**	Homonym
	2	-thelelo isithelelo izithelelo	species herb	
	3	-sithelo isisithelo izisithelo	screen	
kokudla	1	**-dla ukudla ukudla-kudla**	**eating corroding cheating food feast plenty food sharp edge cutting edge instrument**	Homonym
	2	kwesokudla	right hand side	

	3	-dlakudla indlakudla izindlakudla	goat present made young mans people sweetheart ante nuptial visit appetizer	
kwikhensa	1	khenka	run trot pace along	No match is made in this case, because the word is a zululised form of cancer. Nevertheless, there is a dictionary entry for cancer (isigaxa).
	2	khenya	rot	
	3	-khenya ubukhenya	hair	
nekhensa	1	-khenekhene ikhenekhene amakhenekhene	exhausted enervated person flirt loitering girl waiting opportunity talking man unreliable person	No match is made in this case, because the word is a zululised form of cancer. Nevertheless, there is a dictionary entry for cancer (isigaxa).
	2	khenkeneneka	stand gazing gaping slightly open	
	3	-khenyekhenye ikhenyekhenye amakhenyekhenye	person walks heavy rolling gait deceitful crafty person	
nezithelo	1	**-thelo isithelo izithelo**	**fruit crop deed product work**	
	2	-sithelo isisithelo izisithelo	screen	
	3	-thelelo isithelelo izithelelo	species herb	
ngokuxhumana	1	shumana	carry matter concert work intimate friendship	No match is made, since the word is formed through vowel coalescence (nga+uku=ngoku), and the verbal extension –ana added to the stem.
	2	xhuxhuma	restless excited struggle vigorously free	
	3	ngokubani		
thola	1	**thola**	**obtain find come pick adopt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	

Table A24.1 CLEF Query C073 Translations

	Query	Translation	Description	Translation
Original English CLEF query	Norwegian referendum on EU		What was the reactions in the rest of Europe to the negative results of the Norwegian referendum in which Norway decided against membership in the European Union (EU).	
Official translation	Ireferendum yamaNorwegian ephathelene ne-EU	*The referendum of Norway that is about the EU*	IYurobhu yonkana yaphatheka kanjani ngemiphumela engemihle yereferendum yaseNorway lapho iNorway yaphikisana nokuba amalunga e-EU (European Union).	The whole of Europe's reactions that arose which was not nice of the referendum of Norway where Norway contested one another and being members of the EU
Mother tongue translation	Isicelo sokuthola umqondo wabo bonke abantu baseNorway nge-EU.	The application to find out the opinion/view of all people/everybody) by Norway on the EU.	Yenzenjani I-Europe emva kwemiphumela engemihle mayelana neNorwaynereferendamu lapho Norway inqume ukungahambelani nokuba yilunga le-European Union.	The happenings of what sort in Europe after that which arose that is not nice that regards the Norwegian referendum where Norway decided to leave and being not members of the European Union.

Table A24.2 CLEF Query C073 Matching

Word in official translation		Dictionary entry	Translations	Comment
EU	1	-thongo ithongo amathongo	ancestral spirit	No match was possible because the word is a proper name.
	2	babela	burn grass	
	3	babela	pungent bitter acrid	
Europe	1	-kopela ukopela okopela	mean underhanded person sneak	No match was possible because the word is a proper name. Interestingly enough the mother tongue translates Europe to the zululised form iYurobhu.
	2	-popo upopo opopo ipopo	pawpaw	

	3	-peni upeni openi	threepenny piece	
European	1	-kopela ukopela okopela	mean underhanded person sneak	No match was possible because the word is a proper name.
	2	-opha umopha	shed blood track trail blood species sea animal species plant remedy hysteria	
	3	-kopelana isikopelana izikopelana	mean underhanded person	
Norway	1	-cwayi ucwayi ocwayi	tobacco	No match was possible because the word is a proper name.
	2	okwayo		
	3	olwayo		
Norwegian	1	gwegwa	hook catch hook entice bring crooked zigzag roundabout way evasive	No match was possible because the word is a proper name.
	2	-dwengana isidwengana izidwengana	small skin blanket castoff useless object garment	
	3	-Lwezi uLwezi oLwezi	lunar month beginning October	
Union	1	-nono inono amanono	neat tidy person animal cleanly person	No match was possible because the word is a proper name.
	2	-noni inoni amanoni	piece fat fat cut meat	
	3	-noni inoni izinoni	sable antelope	
amalunga	1	-lunga ilunga amalunga	black brown beast white stripes species shrike old person	A homonym that could easily be mistranslated to "joint of the knuckle".
	2	malunga	line opposite anent reference	
	3	**-lungu ilungu amalungu**	**joint knuckle joint internode member society**	
engemihle	1	mihla-ngemihla	day	This word is not matched to a correct dictionary entry, since the stem –imuhle changes through the negative form -nga- and also through vowel coalescence (a+i=e).
	2	ngesihle	freely	

	3	-hlenhle umhlenhle imihlenhle	chasm fissure	
ephathelene	1	phathelela	grip tightly hold lay hands make constant reference	
	2	**phathela**	**handle carry treat mention deputize act behalf act regent**	
	3	phathelana	handle carry treat mention deputize act behalf act regent	
ireferendum	1	-phendumpendu impendumpendu izimpendumpen du	repeated turning backward forward revolving	The word is not matched to a dictionary entry, since it is a borrowed word.
	2	-dumedume udumedume izindumendume	swollen wound	
	3	-fefe ifefe amafefe	blue jay roller bird lilac breasted species Coracias garrulus	
kanjani	1	kanjani	manner	This match must be seen in context with the paraphrase "the sort of after effect" that translates to the reactions of the people.
	2	**-njani**	**sort**	
	3	-njani ubunjani	condition state health improvement health	
ngemiphumela	1	**-phumela umphumela imiphumela**	**small shrub used love charm after effect**	This match must be seen in context with the paraphrase "the sort of after effect" that translates to the reactions of the people
	2	phumela	come rise leave discharge relieve nature	
	3	-nomphumela unomphumela onomphumela	species bushveld shrub bearing fruit Gardenia cornuta	
yaphatheka	1	phaphatheka	lose head excited act speak nervously rush away run wildly stampede fade lose colour	The word would not match, since a verbal extension (-eka) is added to the stem –phatha.
	2	-phaphatheka iphaphatheka amaphaphatheka	nervous excitable person timid person small animal	

	3	-phaphatheka umphaphatheka imiphaphatheka	faded discoloured object	
yaphikisana	1	**phikisana**	**contest dispute compete vie another**	
	2	-phikisani umphikisani abaphikisani	contest dispute compete vie another	
	3	-phikisano impikisano izimpikisano	competition rivalry race bone contention disputation	
yereferendum	1	-yendu isiyendu iziyendu	leaning toppling object	Once again, a match would not be made to the borrowed word.
	2	-phendumpendu impendumpendu izimpendumpendu	repeated turning backward forward revolving	
	3	-dumedume udumedume izindumendume	swollen wound	

Table A25.1 CLEF Query C074 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Inauguration of Channel tunnel***		Find documents describing the inauguration of the channel tunnel and naming the national representatives of Britain and France present at this ceremony	
Official translation	Ukuvulwa kwethonela ngaphansi kwamanzi emkhathini weBritain neFrance	To open the tunnel that is beneath the water that is between Britain and France	Thola imibhalo echaza ngokuvulwa kwethonela ngaphansi kwamanzi emkhathini weBritain neFulansi kanye nokubizwa kwamagama alabo ababemele isizwe saseBritain neFulansi kulo mgubho.	Find documents that explain the opening of the tunnel that is beneath the water that is between Britain and France and give the names that are theirs that represent the nation of Britain and France
Mother tongue translation	Ukugcotshwa komsele kwethonela	To anoint the channel tunnel	Thola izincwadi ezichaza ngokugcotshwa kwetonela komsele kanye nokwethulwa kwalabo abamele iFrance neBritain abazobe bekhona kulomcimbi	Find documents that explain the anointment of the channel tunnel and that which is quiet that stand for France and Britain that would be present at the engagement

Table A25.2 CLEF Query C074 Matching

Word in official translation		Dictionary entry	Translations	Comment
Britain	1	-thinta ithinta amathinta	grave sepulchre tomb	No match was possible because the word is a proper name.
	2	titinya	examine closely feel carefully inquire	
	3	kithina		
France	1	-ncengancenga incengancenga izincengancenga	fretful peevish child person always needs humouring	No match was possible because the word is a proper name. A zululised form eFulansi does exist that is found in the dictionary.
	2	nce	tinkling metallic sound ticking	

	3	nce	order	
ababemele	1	-meleli ummeleli abameleli	advocate	The word would not match to any of the dictionary entries, since it is formed through the addition of a verbal extension (-ele) to the stem –babema.
	2	-bambelelo ibambelelo amabambelelo	verandah pole	
	3	-bemi umbemi ababemi	smoker takes snuff man striking personality	
alabo	1	labo		No match could be found because the word is a demonstrative in the second position of Class 2 - the word itself has no meaning.
	2	-hlabo isihlabo izihlabo	small piece wood scraping skin fork	
	3	-hlabo uhlabo izinhlabo	small piece wood iron points nails knocked scraping skins file	
emkhathini	1	**-khathi umkhathi imikhathi**	**intervening two things intervening distance space earth heaven firmament duration space specified point interval space horizon**	
	2	-khethini ikhetho amakhetho	bridegrooms party wedding dance performed bridegrooms party leaf	
	3	khathi		
isizwe	1	**-zwe isizwe izizwe**	**nation tribe clan state rapidly spreading brain disease mania**	A homonym that could easily have translated to "rapidly spreading brain disease".
	2	-shisizwe ushisizwe oshisizwe	species spreading plant used love-charm	
	3	-shisizwe ushisizwe oshisizwe	species spreading plant used love-charm	
kulo	1	-sukulo umsukulo imisukulo	medicines burnt doctor crops	No match could be found because the word is a demonstrative in the first position of Class 1 - the word itself has no meaning.
	2	-fukulo umfukulo imifukulo	spinal region back	

	3	-cakulo isicakulo izicakulo	ladle	
kwamagama	1	**-gama igama amagama**	**name appellation word letter alphabet statement song hymn air words song notability notoriety**	
	2	-gamu igamu amagamu	name word single part of speech letter alphabet statement song hymn air notability notoriety	
	3	-gaga igaga amagaga	thorax mere skeleton bones emaciated person	
kwamanzi	1	**-nzi amanzi**	**water term applied beer tea beverage unreliable talk weak mindedness poor opinionised reference fear euphemistic term urine**	A homonym that could have several different meanings.
	2	-khamanzi inkamanzi izinkamanzi	lower lip animals	
	3	-khamansi ikhamanzi amakhamanzi	water ladle water dipper	
kwethonela	1	-honela uhonela ohonela	brick kiln	The word would not have a match, since it is a zululised form of tunnel.
	2	onela	spoil injure sin	
	3	-thonyela ithonyela amathonyela	soft pulpy object	
mgubho	1	-gubo umgubo imigubo	face powder Natives	The word forms part of a paraphrased form that means "the unveiling of".
	2	-gubudo umgubudo imigubudo	act covering turning upside down deceit trickery	
	3	**-gubu umgubu imigubu**	**covering veiling protecting**	
ngaphansi	1	**ngaphansi**	**beneath side amount size**	This is another good example of the cultural context in

	2	ngaphane	necessity right	which the word channel tunnel is literally translated to the paraphrase "the channel that is beneath the water".
	3	ngapha	side	
nokubizwa	1	nokuba		This word does not match any of the dictionary entries, since it is formed through palatalisation of the stem ukuba.
	2	ukuzizwa	perceive various senses hear listen give ear obey perceive understand taste smell feel touch sense feel live alive sound good condition feel self important	
	3	-zwakubi inzwakubi izinzwakubi	irascible impetuous person hasty temper inconsiderate action	
thola	1	**thola**	**obtain find come pick adopt receive family**	
	2	-thola ithola amathola	mesh spiders web	
	3	-thola intola izintola	find lucky strike scarcity meagre picking scantiness	
ukuvulwa	1	-zukulwane isizukulwane izizukulwane	offspring generation descendant	The word would not match because of palatalisation of the verb stem -vula to –vulwa. Interestingly, the equivalent mother translation of the word ukuvulwa is ukugcotshwa. It appears that a totally incorrect translation were made, but when put in a cultural context, it is clear that the translator did not take the context of the sentence into consideration, but rather the word itself. For the Zulus, an inauguration is normally associated with royalty. Only kings and queens are inaugurated, thus being anointed as holy ones. For them it might be very confusing to actually anoint a tunnel, but it remains the translation of inauguration.
	2	-zukulwane umzukulwane abazukulwane	grandchild	
	3	-vila ubuvula	laziness	

Table A26.1 CLEF Query C075 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Euskirchen court massacre***		Find documents on the courthouse massacre in Euskirchen, Germany, in which 7 people died.	
Official translation	Ukubulawa ngesihlungu enkantolo e-Euskirchen	To be in a lying position that is poisonous at the court of Euskirchen	Thola imibhalo ngokubulawa ngesihlungu enkantolo yase-Euskirchen, eJalimane, lapho kwashona khona abantu abayisikhombisa	Find documents of those in a lying position that is poisonous in the court of Euskirchen, of Germany, where there died people that is seven
Mother tongue translation	Umbhicongo wasenkantolo yase-Euskirchen	The destruction of the court of Euskirchen	Thola izincwadi ezikhuluma ngombhicongo wasenkantolo yase Euskirchen, eJalimane, lapho kwafa khona abantu abayisikhombisa	Find documents that speak of the destruction of the court of Eurskirchen, in Germany, where died there people that is seven

Table A26.2 CLEF Query C075 Matching

Word in official translation		Dictionary entry	Translations	Comment
Euskirchen	1	-chene isichene izichene	loin skin	No match was made because the word is a proper name, taking only a class prefix.
	2	-cheni umcheni abacheni	bridegroom	
	3	-chene isichene izichene	loin skin	
Germany	1	-manya umanya izimanya	smile face beautiful women beauty beautiful woman	No match was made because the word is a proper name, taking only a class prefix. There appears to be a zululised form eJalimane though.
	2	-nya inya amanya	desire exert oneself order make loss feeling vengeance vindictiveness castoff meltings belly snake private unpleasant things	

	3	-manyanya umanyanya omanyanya	species forest shrub Grumilea capensis	
abantu	1	**-ntu umuntu abantu**	**human being person man member African native race black man person human feelings subject servant dependant**	
	2	bantu	people say wonderful dear believe	
	3	-nta- umnta- abanta	child	
abayisikhombisa	1	**-khombisa isikhombisa izikhombisa**	**index finger forefinger pointer indicator seven**	
	2	khombisa	cause help help point point show	
	3	-khombe isikhombe izikhombe	species thorn shrub mistbelt area Acacia natalitia	
bashona	1	**shona**	**sink sight disappear set sun moon die lose heavily poor bankrupt ruined**	Homonym
	2	shobashoba	wriggle wriggle move restlessly fidget	
	3	-shobashoba isishobashoba izishobashoba	restless person	
enkantolo	1	**-kantolo inkantolo izinkantolo**	**magistrates court charge office**	
	2	-santolo isisantolo izisantolo sántu	species of shrubby climber tough thing easily torn broken	
	3	-ntolo intolo intol'ebomvu	weakness knees species shrub	
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	

	2	**-balo umbalo imibalo**	**document writing entry mark**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	
ngesihlungu	1	**-hlungu isihlungu izihlungu**	**poison snake venom antidote poison nettle rash medical treatment ensure venomous results fights strikes another**	
	2	-hlungu ihlungu amahlungu	newly burnt veld burnt grass patch	
	3	ngesihle	freely free	
ukubulalwa	1	**bulalwa**	**lying**	
	2	-lala ubulala	qualities Lala people	
	3	vukubula	remove lid take covering turn mouth path	

Table A27.1 CLEF Query C076 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	Solar energy		In what applications is solar energy being used or being considered for future use?	
Official translation	Amandla adonswa elangeni	*Power that is of the planet of the sun*	Amandla adonswa elangeni angasetshenziswa kanjani noma angasetshenziselwani esikhathini esizayo?	Power from the planet of the sun causes or helps in the future assistance?
Mother tongue translation	Amandla elanga	*Power of the sun*	Kucatshangwa ukuthi amandla elanga angasetshenziswa kanjani noma angasetshenziselwani esikhathini esizayo?	Think of things where power of the sun causes or helps in assisting the time that is to come

Table A27.2 CLEF Query C076 Matching

Word in official translation		Dictionary entry	Translations	Comment
adonswa	1	donsa	pull draw drag tuck strain attract allure entice steep absorb	In this instance the match is made to the homonym, and actually should be read in context with the paraphrase "the planet of the sun",
	2	**-donsa indonsa**	**planet Jupiter planet Mars**	
	3	-donsa isidonsa izidonsa	griping straining stool	
amandla	1	**-andla amandla**	**strength power moral strength authority ability**	
	2	-ndla amandla	strength	
	3	-mandla	mighty strong powerful	
angasetshenziselwani	1	ngasenzansi	lower part	
	2	enzisela	cause make help calf new mother	
	3	hlanganisela	assemble join make dense ward add surround mix together collate compile	
asetshenziselwani	1	**enzisela**	**cause make help calf new mother**	The match is made to "help" that translates to assistance. Note the palatalisation (b>tsh) that changes the stem.
	2	ntshentshetha	trot short quick steps	
	3	-maselwane umaselwane omaselwane	species wild cucumber	
elangeni	1	-phephelelangeni umphephelelangeni imiphephelelangeni	assegai wood tree	No match is made in this case, since the stem –langa changes through the locative forming of a>eni.
	2	-Mpangeni iMpangeni	place mouth Umhlathuze river	
	3	-magangeni umagangeni omagangeni	roamer hills	
esikhathini	1	**-khathi isikhathi izikhathi**	**time clock watch timepiece**	This translation of this word is actually paraphrased to "the time that is still to come".

	2	-khathi ikhathi amakhathi	opportunity chance time	
	3	-khatha isikhatha izikhatha	rolled object entangled ball knot people hair ball found stomach calves	
esizayo	1	-ayoyo isayoyo izayoyo	new born young man animal bird	
	2	**siza**	**help aid assist succour**	
	3	-siza isiza amasiza	bluish coloured sandstone	

Table A28.1 CLEF Query C077 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	***Teenage suicides***		What information is available concerning teenage suicides?	
Official translation	Ukuzibulala kwabantu abasha	To kill a person that is young	Yimiphi imibiko yolwazi ekhona mayelana nokuzibulala kwabantu abasha?	What reports are there that exist regarding the killing of people that is young?
Mother tongue translation	Ukuzibulala kwentsha	To kill those that is in their youth	Kukhona luphi ulwazi mayelana nokuzibulala kwentsha na?	What information exists regarding the killing of those in their youth?

Table A28.2 CLEF Query C077 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
abasha	1	**-sha umusha abasha**	**sweetheart young bride**	
	2	basha	walking dragging limb flat feet	
	3	shabashaba	wriggle fidget move restlessly	
ekhona	1	Hhayi-bo Hhayi khona Hhayikhona	stop shame	

	2	**khona ukubakhona ukubakho**	**present presence existence**	
	3	khona	order	
kwabantu	1	**-ntu umuntu abantu**	**human being person man member African Native race black man person human feelings subject servant dependant**	Homonym
	2	-ntakwabo umntakwabo abantakwabo	member mothers hut blood relation sweetheart	
	3	kwalahlabantu	place execution	
nokuzibulala	1	bulala	lying position recumbent posture	The homonym could also have matched the first dictionary entry, resulting in a mistranslation.
	2	**bulala**	**kill murder put death destroy break pieces hurt harm injure ill treat poisonous lay heavy weight persons body adduce evidence convict weak evidence**	
	3	-bulalambiza isibulalambiza izibulalambiza	species small bird	
ukuzibulala	1	-lala ubulala	quality Lala people	
	2	bulala	lying position recumbent posture	
	3	**bulala**	**kill murder put death destroy break pieces hurt harm injure ill treat poisonous lay heavy weight persons body adduce evidence convict weak evidence**	
yimiphi	1	-phini umphini imiphi	haft handle axe	There is no match for this word, that has the stem -phi, as a result of object concord (yi-) that is added to the Class prefix (imi-).
	2	-pha umupha imipha	maize stalk cob	
	3	-phimbi umphimbi imiphimbi	Lowveld fruit tree	
yolwazi	1	-lolwazi umlolwazi imilolwazi	whetstone	No match, since the word is formed through vowel coalescence (ya+u=yo).

	2	-klwazi inklwazi	rippling babbling sound sound womens voices singing	
	3	-hlwazi umhlwazi imihlwazi	species tree rare protea species green striped snake	

Table A29.1 CLEF Query C078 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Venice film festival***		Which film or films won the Golden Lion in the 51st Venice Film festival in September 1994?	
Official translation	I"Film Festival" laseVenice	The film festival of Venice	Yiliphi ifilimu noma amafilimu awine iGolden Lion ye"51st Film Festival" laseVenice ngoSepthemba ka-1994?	Which film or films won the Golden Lion at the 51st Film festival of Venice in September 1994?
Mother tongue translation	Umgubho wombukiso waseVenice	The festival of films of Venice	Yisiphi isithombe noma yiziphi izithombe ezazuza iGolden Lion kumbukiso wamashumi amahlanu nanye waseVenice ngoSepthemba ka-1994?	Which picture or which pictures (films) obtained the Golden Lion at the festival of films that is 51st of Venice in September 1994?

Table A29.2 CLEF Query C078 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
1994	1	-thongo ithongo amathongo	ancestral spirit	There is no match in this case, since the numeral is preceded by an inflected prefix.
	2	babela	burn grass	
	3	babela	pungent bitter acrid	
51st	1	st	digraph	There is no match in this case, since the numeral is preceded by an inflected prefix.
	2	-krestu ikrestu amakrestu	Christian	
	3	-krestu ubukrestu	Christian behaviour life	

Festival	1	etiva	behind	This word forms part of a borrowed word paraphrased in "51st Film Festival".
	2	-valo isivalo izivalo	door cork stopper lid shutter astringent medicine stopping diarrhoea	
	3	-vesiti ivesiti amavesiti	vest	
Film	1	-lima isilima izilma	deformed person beast thing fool idiot	This word forms part of a borrowed word paraphrased in "51st Film Festival".
	2	-lima ilima amalima	communal cultivation working bee cultivating field ploughed	
	3	-lomo ilomo amalomo	word words	
Golden	1	olwenu		This is a borrowed word that only takes a Class prefix, and therefore would not match to any dictionary entry.
	2	-kolweni ukolweni	corn wheat brown bread	
	3	-golweni ugolweni ogolweni	five shillings	
Lion	1	-nono inono amanono	neat tidy person animal cleanly person	This is a borrowed word that only takes a Class prefix, and therefore would not match to any dictionary entry.
	2	-ono umono imono	crate carrying fowls wicker trap catching birds	
	3	-oni isoni izoni	sinner habitual wrong doer criminal	
Septhemba	1	themba	hope trust expect faith put confidence rely	This is a zululised word and does not match to any dictionary entry.
	2	-themba ithemba amathemba	trust hope faith expectation reliance belief	
	3	-themba intemba	trust reliance expectation	
Venice	1	engce	outside	This is a borrowed word that only takes a Class prefix, and therefore would not match to any dictionary entry.
	2	-yence iyence amayence	wedding wedding party	
	3	-ngcengce ingcengce izingcengce	large basket	

amafilimu	1	**-filimu ifilimu amafilimu**	**film photographic film plate bioscope show**	Although this word is a zululised form, it does have a dictionary entry.
	2	-fidi ifidi amafidi	foot measure length	
	3	-limi amalimi	impediment speech stuttering stammering	
awine	1	-phawini uphawini ophawini	share apportionment ration	This homonym could have easily been mistranslated to "ricksha boy"
	2	-thawini ithawini amathawini	town	
	3	**-wini umwini abawini**	**winner ricksha boy**	
ifilimu	1	**-filimu ifilimu amafilimu**	**film photographic film plate bioscope show**	Although this word is a zululised form, it does have a dictionary entry.
	2	-ncifili incifili amancifili	expression dislike contempt	
	3	-lima ilima amalima	communal cultivation working bee cultivating	
noma	1	noma		No match for this conjunction (na+uma) meaning "whether...or".
	2	oma	dry dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
yiliphi	1	-philiphili iphiliphili amaphiliphili	headstrong person	There is no match for this word, that has the stem –phi, as a result of object concord (yi-) that is added to the Class prefix (imi-).
	2	-hliphisi ihliphisi amahliphisi	slipper	

	3	-hliphihliphi amahliphihliphi	disorderly state	

Table A30.1 CLEF Query C079 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Ulysses space probe***		Find documents that describe the European space probe mission Ulysses or discuss its objectives	
Official translation	Ukuhlolisiswa komkhathi i-Ulysses	To explore of space by Ulysses	Thola imibhalo echaza ukuhlolisiswa komkhathi I-Ulysses lwaseYurobhu noma uxoxe ngezinjongo zako.	Find documents that explain the exploration of space by Ullysses in Europe or give an account of the aim of it.
Mother tongue translation	Indiza yasemkhathini yaseUlysses	The aeroplane of space of Ulysses	Thola umbhalo ocacisa kabanzi ngenhlosa yendiza yasemkhathini yase-Ulysses, noma chaza ngokusophwe yilendiza	Find documents that explain the intention of the aeroplane of Ullysses, or explain the determination/purpose of the aeroplane

Table A30.2 CLEF Query C079 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
Europe	1	-kopela ukopela okopela	mean underhanded person sneak	This is a borrowed word that only takes a Class prefix, and therefore would not match to any dictionary entry. However, there is a dictionary entry for iYurobhu.
	2	-popo upopo opopo ipopo	pawpaw	
	3	-peni upeni openi	threepenny piece	
Ulysses	1	-sese isese	secret place place hiding	This is a borrowed word that only takes a Class prefix, and therefore would not match to any dictionary entry.
	2	-risese urisese orisese	recess interval school lessons break	

	3	seseza	whisper spread confidential information lead gently coax	
echaza	1	**chaza**	**incisions tattoo incise order rub medicine reveal explain show meaning dominate crackle**	
	2	-chaza imichaza	greens	
	3	chachaza	dominate lord assume unrightful authority rude act respect drip fall drops crackle	
komkhathi	1	**-khathi umkhathi imikhathi eng**	**space intervening two things intervening distance space earth heaven firmament duration space time specified position horizon**	
	2	khathi	when	
	3	-khathi ikhathi amakhathi	opportunity chance time doing something	
ngezinjongo	1	**-jongo injongo izinjongo**	**aim purpose object**	
	2	-njongoni injongoni izinjongoni	wildebeest	
	3	-jongo injongo izinjongo	aim purpose object	
noma	1	noma		No match for this conjunction (na+uma) meaning "whether...or".
	2	oma	dry dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
ukuhlolisiswa	1	**hlolisa**	**cause live ease attack surprise ambush**	
	2	-wulukuhlu isiwulukuhlu iziwulukuhlu	poring bulk mass expulsion abnormally bulky object	

	3	-hholisaka uhholisaka ohholisaka	large sack	
uxoxa	1	**xoxa**	**narrate tell give account hold conversation converse chat court woo attempt favour**	
	2	xoxoma	hop bob restless unsettled	
	3	-xoxo ixoxo amaxoxo	narrative account pleasant talk conversation chat frog	
zakho	1	zakho		No match, since the word is the negative form of the adverbial form -khona, with the appropriate class prefix added.
	2	ezakho		
	3	akho		

Table A31.1 CLEF Query C081 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	IRA attacks in airports		Find documents that describe terrorist acts by the Irish Republican Army (IRA) in European airports	
Official translation	Ukuhlaselwa kwe-IRA ezikhumulweni zezindiza	The attack by the IRA of the place of airplanes	Thola imibhalo echaza ngezenzo zobuphekulazikhuni be-IRA ezikhumulweni zezindiza zaseYurobhu	Find documents that report on acts of militants of the IRA of the place of airplanes in Europe
Mother tongue translation	Ukuhlasela kwe-IRA ezikhumulweni zezindiza	The attack of the IRA of the place of airplanes	Thola imibhalo echaza kabanzi ngezenzo zobuqhwaga ze-Irish Republican Army (IRA) ezikhumulweni zezindiza zaseEuropa	Find documents that report on acts of taking by force by the IRA of the place of airplanes in Europe

Table A31.2 CLEF Query C081 Matching

Word in official translation		Dictionary entry	Translations	Comment
IRA	1	-reki ireki amareki	rake	No match would be made to this proper name. It only takes a class prefix.
	2	-roza iroza amaroza	rose rose bush	
	3	-rula irula amarula	ruler	
Islam	1	lamu	stop stay peace	No match would be made to this proper name. It only takes a class prefix.
	2	-lamu umlamu abalamu	brother law sister law	
	3	lami		
ekuthathweni	1	-thathawe uthathawe izintathawe	species straggling thorn bush stiff backed round shouldered person person walks stoop	No match was made to the stem (-thathe) because of locative forming (e>weni).
	2	-thathe uthathe izintathe	sneezewood tree undergrowth dry grass	
	3	-khathwa ukhathwa okhathwa	medicinal plant used hysteria	
ethathwe	1	thetha	scold find fault noisily nag try law case preside lawsuit offer praises prayer sacrifices spirits judgement favour find guilty sympathy attracted	
	2	-thathawe uthathawe izintathawe	species straggling thorny bush stiff backed round shouldered person person walks stoop	
	3	**-thathe**	**act suddenly without precaution preparation**	
iqembu	1	**-qembu iqembu amaqembu**	**detached portion division detachment small group herd**	

	2	-qembuqembu ubuqembuqembu	disagreement division opinion	
	3	qembuka	divided separated divide two parts detached	
kwendiza	1	ndiza	fly excited lose mental balance excitement	A match is made, but it must be noted that word forms part of the paraphrase "the place of airplanes" that translates to airport.
	2	**-ndiza indiza izindiza**	**aeroplane**	
	3	-ndizane indizane izindizane	vascillating person easily influenced swings opinion bird fly	
lezimpi	1	**-mpi impi izimpi**	**regiment army military force encounter fight engagement battle war hostile person band foe**	
	2	imp- izimp	nouns commencing	
	3	-phici uphici izimpici	bleary eyed person person white spot eye	
nendikimba	1	-dikimba indikimba	bulk whole main	No match, because of pre-nasalisation (b>mb) of the stem (-ndikiba).
	2	-ndikimana amandikimana	heavy squat-bodied vessel broad flat base	
	3	-kimba umkimba imikimba	wedding party	
ngendluzula	1	-dluzula indluzula izindluzula	violence wrench force	
	2	**dluzula**	**snap jerk tug pull violently act violence**	
	3	-dluzula indluzula izindluzula	violence wrench force	
ukuthathwa	1	-khathwa ukhathwa okhathwa	medicinal plant hysteria	No match was made because of pre-nasalisation (-w-) of the stem.

	2	tha yitha ukutha	give name name nickname pour vessel small aperture pour funnel inject enema select pick best	
	3	-tha ukutha	custom referring person impersonal concords	
yonke	1	yonke		All of these words are inclusive quantifiers in their respective classes, and does not have any meaning.
	2	onke		
	3	konke		

Table A32.1 CLEF Query C084 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Shark attacks***		Documents will report any information relating to shark attacks on humans	
Official translation	Ukuhlasela umfingo	To attack by sharks	Imibhalo iyobika nganoma yiluphi ulwazi oluphathelene nokuhlasela kwabantu umfingo	Documents will report on information in any direction that mentions the attack of sharks on people
Mother tongue translation	Ukuhlaselwa oshaka	To attack by sharks	Imibhalo izokhipa umkhondo ophathelene nokuhlaselwa kwabantu ngoshaka	Documents will draw a thorough understanding that is there that mentions the attack on people by sharks.

Table A32.2 CLEF Query C084 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
imibhalo	1	-hhalo umhhalo imihhalo	harrowed field	
	2	**-balo umbalo imibalo**	**mark document entry writing scriptures large blanket**	
	3	-phalo umphalo imiphalo	scraped hide prepared skin deep hole	

iyobika	1	-yobinga iyobinga amayobinga	tuft grass hair	
	2	-yobi isiyobi iziyobi	tuft grass bushes left after cutting	
	3	**bika -vika**	**report introduce announce warn omen**	
koshaka	1	-shakashaka umshakashaka imishakashaka	species Panicle millet	The word is not match to any of the (correct) dictionary entries, since it's a zululised form of shark.
	2	shaka	dish	
	3	-shaka ishaka amashaka	gripes stomach ache early pregnancy small beetle swallowed cure bed wetting	
ngaluphi	1	**ngalapha**	**side direction**	
	2	ngalapho	direction side	
	3	-luphici	bleary eyed white spot eye	
nganoma	1	-sangano insangano izisangano	confused state mind	No match for this conjunction (na+uma) meaning "whether...or".
	2	noma		
	3	-gano umgano imigano	bullock	
nokuhlasela	1	hlasela		
	2	-hlaseli umhlaseli abahlaseli	attacker invader hunter	
	3	**-hlaselo inhlaselo izinhlaselo**	**attack invasion hunt**	
oluphathelene	1	**phathelela**	**grip tightly hold lay hands make consistent reference quote authority**	
	2	-luphatha		

	3	phathelana	handle carry treat mention deputize act behalf act regent	
ulwazi	1	**-azi ulwazi**	**thorough knowledge understanding**	
	2	-klwazi inklwazi	rippling babbling sound sound women singing	
	3	-hlwazi umhlwazi imihlwazi	species rare tree protea veld species green striped snake	

Table A33.1 CLEF Query C086 Translations

	Query	**Translation**	**Description**	**Translation**
Original English CLEF query	Renewable power		Find documents describing the use of, or policies regarding "green" power, i.e. power generated from renewable energy	
Official translation	Amandla angavuselelwa	Power that is renewable	Thola imibhalo echaza ngokusetshenziswa noma imigomo maqondana ne"green power" (i"green power" itholwa emandleni angavuselelwa).	Find documents that show the meaning of the help of or the purpose that refers to "green power" ("green power" that is obtained from power that is renewable).
Mother tongue translation	Amandla angenziwa kabusha	Power that can be made new	Thola umqulu ochaza ukusetshenziswa noma umgomo mayelana namandla aluhlaza (umzekelo amandla avezwa umdlandla owenziwe kabusha).	Find information that joins together work or the purpose regarding power that is green that relates to power that brings forth the keenness to make new

Table A33.2 CLEF Query C086 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
amandla	1	**-andla amandla**	**strength power moral strength authority ability**	
	2	-ndla amandla	strength	

	3	-mandla	mighty strong powerful	
angavuselelwa	1	**vuselela**	**renew repair renovate revive rebuild remind call memory**	In this case the word is matched, although it appears in the palatalised form where a passive -w- is added to end.
	2	-vuselelo imvuselelo izimvuselelo	reminder warning renovation renewal repair revival	
	3	-mangalelwa ummangalelwa abamangalelwa	defendant case	
emandleni	1	-phandleni impandleni	large white yellow beads	No match was made, since the stem (-mandla) changed to –mandleni through locative forming (a>eni).
	2	-Mpandleni	place name	
	3	-andle ulwandle	sea	
green	1	gwee	quacking ducks	There is no match for this borrowed word. Interestingly enough though, the mother tongue translation attempts to match the word to the color green. In this case however, the word is part of the paraphrased "green power" that translates to environment-friendly renewable power.
	2	-nengenenge unengenenge izinengenenge	endless drawn matter continuous following persecution	
	3	-gengenene amagengenene	wide spreading horns furrow corrugation rams horns	
imigomo	1	**-gomo umgomo imigomo**	**essence true facts real truth standard limit strict custom transgressed goal purpose main object aimed**	
	2	-gogo umgogo imigogo	emaciated person animal shriveled dried carcass skeleton field	
	3	-goyo umgoyo imigoyo	retirement seclusion	
itholwa	1	**-tholwa isitholwa izitholwa**	**adopted child adopted family**	This example illustrates the cultural context, where the translation is matched to "adopted". An attempt is made whereby it is illustrated that the renewable power is adopted from green power – as one would adopt a child.
	2	-thola ithola amathola	mesh spiders web	

	3	-kholwa ikholwa amakholwa	believer Christian	
maqondana	1	**maqondana**	**line straight direction opposite reference appertaining**	
	2	qondana	understand opposite line	
	3	-qondo iqondo amaqondo	stitch single binding Native sewing directions administering medicine stone bladder	
ne	1	ne	sticking tight adhering suffocation enclosing	In this case the incorrect match is made various entries for "ne". In this example, however, the word is a class prefix that is added to borrowed word, but changes through vowel coalescence.
	2	-ne	four	
	3	-ne isine izine	fourth place fourth	
ngokusetshenziswa	1	ngokushesha	quickly hurriedly	A good example of palatalisation (b>tsh) of the verb stem sebenza.
	2	**enziswa**	**cause help make**	
	3	tshengisa	cause barter sell	
noma	1	noma		No match for this conjunction (na+uma) meaning "whether...or".
	2	oma	dry dry thirsty thin wiry	
	3	ngoma	sing dance song sing hymns	
power	1	-pokwe impokwe izimpokwe	grass flower	This borrowed word is not matched, since it forms part of the paraphrased "green power" and only takes an appropriate class prefix.
	2	-jeri ujeri ojeri	Jack fruit tree	
	3	-ampokwe isampokwe izampokwe	sjambok thick strip hippopotamus hide whip	

Table A34.1 CLEF Query C088 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Mad cow disease in Europe***		Find documents that cite cases of Bovine Spongiform Encephalopathy (the mad cow disease) in Europe	
Official translation	I-"Mad Cow Disease" eYurobhu	The Mad Cow disease in Europe	Thola imibhalo ebika ngezilwane ezine"Bovine Spongiform Encephalopathy" (mad cow disease) eYurobhu	Find documents that report on animals with mad cow disease in Europe.
Mother tongue translation	Isifo samatele e-Europa	The disease of the hoofs of animals in Europe	Thola umqulu oveza izigameko zesifo samatele e-Europe	Find information that brings forth matters of sickness of the hoofs of animals in Europe

Table A34.2 CLEF Query C088 Matching

Word in official translation		Dictionary entry	Translations	Comment
Bovine	1	-hhovini uhhovini ohhovini	oven	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Bovine Spongiform Encephalopathy".
	2	yomine	four four	
	3	zozine	four four	
Cow	1	-wowo iwowo amawowo	large heap	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Mad Cow Disease".
	2	lowá		
	3	wowu	fine	
Disease	1	-sense isense amasense	chapping feet	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Mad Cow Disease".
	2	-sese isese	secret place place hiding	
	3	-sasa isasa amasasa	morning breakfast morning meal	

Encephalopathy	1	-phalo iphalo amaphalo	scraping hide piece skin scraped dressing hide dried emaciated body	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Bovine Spongiform Encephalopathy".
	2	-phalo impalo izimpalo	scraping hide deep hole chasm ravine	
	3	-phalo isiphalo iziphalo	iron scraper softening hide scraping mud metal identification disk worn labourers	
Mad	1	dada	puzzle cause person tobe loss place difficult circumstances	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Mad Cow Disease".
	2	-dada idada amadada	duck cultivated edible tuber	
	3	-dada isidada izidada	expansive garden	
Spongiform	1	-ongi umongi abongi	nurse hospital nurse	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Bovine Spongiform Encephalopathy".
	2	-fongofongo umfongofongo imifongofongo	species tree	
	3	-fongafonga umfongafonga imifongafonga	common tree hygrophilous coast bush	
cow	1	co	falling drop liquid regular intervals	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Mad Cow Disease".
	2	-nco	speckled red white	
	3	-nco inco izinco	beast speckled red white species speckled bird	
disease	1	-sense isense amasense	chapping feet	
	2	-sese isese	secret place place hiding	
	3	-sasa isasa amasasa	morning morning meal breakfast	
ebika	1	**bika -vika**	**report introduce warn omen**	
	2	-bika ibika amabika	harbinger omen species ant white spots	
	3	-bibika imbibika izimbibika	cry baby mournful faced person	

ezine	1	**ezin-**	**concord class**	
	2	-ne isine izine	fourth place fourth	
	3	-nezo inezo izinezo	blanket	
i	1	-bemakancane i bemakancane amabemakancane	moderate smoker snuff taker	This would not match to any dictionary entry, since it has not meaning and serves as a class prefix.
	2	-xhama i xhama amaxhama	womans fibre grass girdle	
	3	-su u lusu izinsu	first stomach cattle paunch human stomach tough pliable object soft goatskin cloak worn women	
mad	1	-madi imadi izimadi	letter	No match is made because the word is a borrowed word and only takes a class prefix. It also forms part of the paraphrased "Mad Cow Disease".
	2	-da amada	water	
	3	ma-	enumerative concord meaning	
ngezilwane	1	**-lwane isilwane izilwane**	**animal animal life wild beast animal person outstanding qualities**	An example of a homonym that could easily have been mistranslated to "a person with outstanding qualities".
	2	-lilwane umlilwane imililwane	small transparent crimson bead crimson coloured object	
	3	-lwanga ulwanga izilwanga	palate roof mouth	

Table A35.1 CLEF Query C090 Translations

	Query	Translation	Description	Translation
Original English CLEF query	***Vegetable exporters***		What countries are exporters of fresh, dried or frozen vegetables?	
Official translation	Abantu abathengisa imifino kumazwe angaphandle	People that sell/trade vegetables from the country to the outside	Imaphi amazwe athengisela ngaphandle imifino emisha eyomile noma eqandisiwe	Which countries are sellers to the outside of vegetables that is fresh and dried or that is very cold like ice?
Mother tongue translation	Imifino ethunyelwa kumazwe angaphandle	Vegetables that go towards the outside.	Imaphi amazwe athumela imifino enothile, eyomisiwe noma eqandisiwe	Which countries send vegetables that are well and fresh and that is very cold like ice?

Table A35.2 CLEF Query C090 Matching

Word in official translation		**Dictionary entry**	**Translations**	**Comment**
abantu	1	**-ntu umuntu abantu**	**man person human member African Native race black man person human feelings subject servant dependant**	
	2	bantu	people say wonderful dear believe	
	3	-nta- umnta- abanta	child	
abathengisela	1	**-thengisi umthengisi abathengisi**	**seller tradesman hawker**	Note the verbal extension (-ela) at the end of the stem (-thengisi).
	2	-thengeli umthengeli abathengi	salesman shop assistant	
	3	-theni umtheni abatheni	castrator animals	
imifino	1	**-fino umfino imfino imifino**	**ordinary food green vegetable edible herb cooked food**	

	2	-fino isifino izifino	daily ordinary food any kind tasty dish European food	
	3	-fingqo umfingqo imifingqo	jerky movement finger joints carerpillar walking	
kumazwe	1	-mazwenda umazwenda omazwenda	species stout forest climber	The word does not match any dictionary entry, since the stem (-zwe) changes because of vowel elision.
	2	umazwenda omnyama	Uvaria caffra Artabotrys monteiroae	
	3	umazwenda omhlophe	Popowia caffra	
angaphandle	1	**ngaphandle**	**externally right independently**	This match actually translates to "sellers to the outside of", which is a paraphrased form of export.
	2	-phandle amaphandle	suburbs places habitation main village outskirts main kraal people live suburbs people acquired local customs uncultured people	
	3	phandle		
imaphi	1	-maphu imaphu izimpahu	map	There is no match, since the word is an enumerative form meaning "which one".
	2	-maphipha umaphipha omaphipha eng	herbalists name tree used enema	
	3	-phiva iphiva amaphiva	black water buck	
amazwe	1	**-zwe izwe amazwe**	**country land people land populace**	
	2	-zwi izwi amazwi	voice order command message word eng twt	
	3	-zwani izwani amazwan	long toe bird	
athengisela	1	**-thengisi umthengisi abathengisi**	**seller tradesman hawker**	

	2	thengisa	cause barter sell	
	3	-thengiso intengiso izintengiso	price commercial dealings trade commerce	
imifino	1	**-fino umfino imfino imifino**	**ordinary food green vegetable edible herb cooked food**	
	2	-fino isifino izifino	daily ordinary food kind tasty dish European food	
	3	-fingqo umfingqo imifingqo	girdle girding high loins	
emisha	1	-mishani imishani izimishani	mission mission station	No match, since the word is a locative formed by the e- being added to the stem -sha.
	2	-shaba umshaba imishaba	temporary hut replacing destroyed grass screen sheltering hut lowest ground layer matting hut keep thatch firm	
eyomile	1	**-omile**	**dry dried thirsty**	
	2	yomine	four four	
	3	-dumile	famed renowned notorious	
eqandisiwe	1	qandisela	bear children rapid succession lay eggs profusely prolific aim ahead moving object provide contingency	The word is not matched to any dictionary entry, since the stem (-qandisa) changes because of a verbal extension (-iwe) being added.
	2	-qandiselo umqandiselo imiqandiselo	provision riddle	
	3	-andisela isandisela izandisela	unnecessary continual addition	

Appendix B

A	an	been	d	everybody
about	and	behind	day	everyone
above	another	before	do	everything
according	any	beforehand	does	excepting
across	anybody	being	doing	everywhere
after	anyhow	below	dost	except
afterwards	anyone	beside	doth	excepted
again	anything	besides	double	exception
against	anyway	between	down	exclude
albeit	anywhere	beyond	dual	excluding
all	apart	both	during	exclusive
almost	are	but	each	far
alone	around	by	either	farther
along	as	can	else	farthest
already	at	cannot	elsewhere	few
also	be	canst	enough	ff
although	became	certain	et	first
always	because	choose	etc	for
among	become	contrariwise	even	formerly
amongst	becomes	cos	ever	forth
am	becoming	could	every	forward

from	hereafter	inasmuch	km	mostly
front	hereby	inc	l	more
further	herein	include	last	mr
furthermore	hereto	included	latter	mrs
furthest	hereupon	including	latterly	ms
g	hers	indeed	less	much
get	herself	indoors	lest	must
go	him	inside	let	my
had	himself	insomuch	li	myself
halves	hindmost	instead	like	namely
hardly	his	into	little	need
has	hither	inward	ltd	neither
hast	hitherto	inwards	many	never
hath	how	is	may	nevertheless
have	however	it	maybe	next
he	howsoever	its	me	nn
hence	i	itself	meantime	no
henceforth	ie	just	meanwhile	nobody
her	if	ke	might	none
here	il	kind	moreover	nonetheless
hereabouts	in	kg	most	noone

nope	ought	sang	slung	staves
nor	our	save	slunk	still
not	ours	saw	smote	such
nothing	ourselves	see	so	supposing
notwithstanding	out	seeing	some	than
now	outside	seem seemed	somebody	that
nowadays	over	seeming	somehow	the
nowhere	own	seems	someone	thee
of	p	seen	something	their
off	per	seldom	sometime	them
often	perhaps	selves	sometimes	themselves
ok	plenty	sent	somewhat	then
on	provide	several	somewhere	thence
once	quite	shalt	spake	thenceforth
one	rather	she	spat	there
only	re	should	spoke	thereabout
onto	really	shown	spoken	thereabouts
or	round	sideways	sprang	thereafter
other	said	since	sprung	thereby
others	sake	slept	st	therefore
otherwise	same	slew	stave	therein

though	upon	whenever	whew	worse
thrice	upward	whensoever	which	worst
through	upwards	where	whichever	thereof
throughout	us	whereabouts	whichsoever	thereon
thru	use	whereafter	while	thereto
thus	used	whereas	whilst	thereupon
thy	using	whereat	whither	these
thyself	very	whereby	who	they
till	via	wherefore	whoa	this
to	vs	wherefrom	whoever	those
together	want	wherein	whole	thou
too	was	whereinto	whom	would
toward	we	whereof	whomever	wow
towards	week	whereon	whomsoever	yet
unable	well	wheresoever	whose	year
under	were	whereto	whosoever	yippee
underneath	what	whereunto	why	you
unless	whatever	whereupon	will	your
unlike	whatsoever	wherever	with	yours
until	when	wherewith	within	yourself
up	whence	whether	without	yourselves

Appendix C

Class no.	Class pref.	Basic pref.	Subject concord	Object concord	Subj. conc. past tense	Rel. conc.	Poss. conc.	Pron. stem	Absol. pronoun	Demonstratives			Quantifiers	
										1	2	3	Incl.	Excl.
1	umu-	m(u)-	u-(e-/a-)	-m(u)-	wa-	o-(a-)	wa+	khe	ye(na)	lo	lowo	lowaya	wonke	yedwa
2	aba-	ba-	ba-(be-)	-b(a)-	ba-	a(ba)-	ba+	bo	bo(na)	laba	labo	labaya	bonke	bodwa
3	umu-	m(u)	u-	-w(u)-	wa-	o-	wa+	wo	wo(na)	lo	lowo	lowaya	wonke	wodwa
4	imi-	mi-	i-	-y(i)-	ya-	e-	ya+	yo	yo(na)	le	leyo	leya	yonke	yodwa
5	i(li)-	li-	li-	-l(i)-	la-	e(li)-	la+	lo	lo(na)	leli	lelo	leliya	lonke	lodwa
6	ama-	ma-	a-(e-)	-w(a)-	a-	a-	a+	wo	wo(na)	la	lawo	lawaya	onke	odwa
7	isi-	si-	si-	-s(i)-	sa-	e(si)-	sa+	so	so(na)	lesi	leso	lesiya	sonke	sodwa
8	izi-	zin-	zi-	-z(i)-	za-	e(zi)-	za+	zo	zo(na)	lezi	lezo	leziya	zonke	zodwa
9	in-	in-	i-	-y(i)-	ya-	e-	ya+	yo	yo(na)	le	leyo	leya	yonke	yodwa
10	izin-	zin-	zi-	-z(i)-	za-	e(zi)-	za+	zo	zo(na)	lezi	lezo	leziya	zonke	zodwa
11	u(lu)-	lu-	lu-	-lu/lw-	lwa-	o(lu)-	lwa+	lo	lo(na)	lolu	lolo	loluya	lonke	lodwa
14	ubu-	bu-	bu-	-b(u)-	ba-	o(bu)-	ba+	bo	bo(na)	lobu	lobo	lobuya	bonke	kodwa
15	uku-	ku-	ku-	-ku/kw-	kwa-	o(ku)-	kwa+	kho	kho(na)	lokhu	lokho	lokhuya	konke	kodwa
16	pha-									lapha	lapho	laphaya		
17	uku-	ku-	ku-	-ku/kw-	kwa-	o(ku)-	kwa+	kho	kho(na)	lokhu	lokho	lokhuya	konke	kodwa
18	mu-													
1 a	u-	m(u)-	u-(e-/a-)	-m(u)-	wa-	o-(a-)	wa+	khe	ye(na)	lo	lowo	lowaya	wonke	yedwa
2 a	o-	ba-	ba-	-b(a)-	ba-	a(ba)-	ba+	bo	bo(na)	laba	labo	labaya	bonke	bodwa
I p.s.		m(u)-	ngi-	-ngi-	nga-	engi-	wa+	mi	mi(na)	lo	lowo	lowaya	wonke	ngedwa
I p.p.		ba-	si-	-s(i)-	sa-	esi-	ba+	ithu	thi(na)	laba	labo	labaya	sonke	sodwa
II p.s.		m(u)-	u-	-ku-	wa-	o-	wa+	kho	we(na)	lo	lowo	lowaya	wonke	wedwa
II p.p.		ba-	ni-	-n(i)-	na-	eni-	ba+	inu	ni(na)	laba	labo	labaya	nonke	nodwa

Note: The brackets show that this section can be elided/fall away under certain circumstances.