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.