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).