Information access in indigenous languages: a case study in Zulu

Erica Cosijn*, Ari Pirkola**, Theo Bothma*, Kalervo Järvelin**
*Department of Information Science, University of Pretoria, Pretoria 0002, South Africa
**Department of Information Studies, University of Tampere, P O Box 607, 33101 Tampere, Finland
ecosijn@postino.up.ac.za, pirkola@tukki.cc.jyu.fi, tbothma@postino.up.ac.za, likaja@uta.fi

Received: 5th September 2002

This study focuses on the intellectual accessibility of information in indigenous languages, using Zulu, one of the main indigenous languages in South Africa, as a test case. Both Cross-Lingual Information Retrieval (CLIR) and metadata are discussed as possible means of facilitating access and a bilateral approach combining these two methods is proposed. Popular CLIR approaches and their resource requirements are analysed and the dictionary-based approach combined with approximate string-matching for query translation from Zulu to English are discussed in detail. Metadata formats for knowledge representation from the Indigenous Knowledge (IK) viewpoint are discussed, in particular the advantages and limitations of the Dublin Core (DC) metadata format.

1. Introduction

Indigenous Knowledge (IK) can be termed as local knowledge, unique to every culture or society. It can also be defined as a large body of knowledge and skills outside the formal education system that enables communities to survive, and is commonly held by communities rather than individuals. IK is tacit knowledge and is therefore difficult to codify, as it is embedded in community practices, relationships and rituals. Recently, there have been several efforts to collect IK in order to analyse it and to prevent it from being lost to posterity. IK has thus been collected into large paper archives and more recently into electronic databases.

Using databases for the representation of IK may offer several advantages. Most importantly, access from a retrieval point of view is much easier in electronic database format than in paper or linear electronic text formats. Secondly, IK can be stored and delivered in multiple copies for those that need it. Furthermore, in database format, it is possible to annotate IK in various ways from multiple viewpoints to facilitate its analysis. However, in order to realise these advantages, IK in databases must be made accessible.

In this study we focus on the problem of intellectual accessibility of IK, as opposed to the problems of database technology (e.g., file structures and their access methods) or availability of equipment. We will analyse and discuss Cross-Lingual Information Retrieval (CLIR) as means of access to such databases. Various problems were experienced, viz. the standard problems associated with CLIR, the lack of suitable electronic resources and the characteristics of Zulu as language. We therefore suggest that one should also consider the use of metadata to describe the content. Regarding CLIR, we analyse popular CLIR approaches and their resource requirements, and discuss in detail the dictionary-based approach for query translation for accessing IK. We also develop means for circumventing the problems caused by insufficient linguistic resources for query translation. Regarding metadata, we analyse metadata formats for knowledge representation from the IK viewpoint and discuss, in particular, the advantages and limitations of the Dublin Core (DC) metadata format.

The paper is organised as follows. In Section 2 we present and analyse basic CLIR approaches and techniques for IK access, and indicate why some of the approaches are not suitable. Section 3 is a discussion of the viability and advantages of using CLIR as access to IK databases in South African indigenous languages (specifically Zulu). In Section 4 we discuss the applicability of metadata to facilitate access to IK databases. Section 5 contains some concluding remarks in which we indicate that retrieval through a combination of CLIR techniques and metadata will enhance precision.

2. The applicability of CLIR for IK databases in a South African context

IK is normally stored in databases in textual form, based on transcribed speech in some indigenous language. Any large IK database is likely to contain IK in several indigenous languages. The analysts and other users of IK may well be capable of reading IK texts directly in several indigenous languages. However, they may have difficulties in expressing their interests properly in these languages - which is the requirement for successful direct content access.

The basic idea of Cross Language Information Retrieval (CLIR) (Hull & Greffenstette, 1996; Oard & Diekema, 1998) is to provide access in one language (the source language) to documents written in another language (the target language). South Africa has a very complex language situation due to the fact that there are eleven official languages, *viz*. English, Afrikaans and nine African languages, including Zulu and Xhosa. In the case of IK, the source language would often be

English or Afrikaans, but could also be an African language. The target language(s) would typically be the indigenous African language(s).

The basic approaches in CLIR involve ***query translation*** from source language to target language(s) and/or ***document translation*** from target language(s) to source language. Document translation requires good machine translation systems for the languages in which the IK documents are written, but such systems will not be available any time soon. Query translation requires fewer resources, but there is the additional requirement that the users should be able to read the IK documents in the languages in which they were written. It is often the case that the user is able to read a foreign language, but is not fluent enough to construct an appropriate query in that language. However, even if the user cannot read the retrieved documents, the user has at least a relevant set of retrieved documents that may be translated manually.

In CLIR, the main strategies for query translation are based on parallel corpora, machine translation and (bilingual) translation dictionaries (Oard & Diekema, 1998). We review below the three methods briefly and then show how the third can be supplemented with approximate string matching.

2.1. Corpus-based

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 (Figure 1). A parallel corpus consists of document pairs such that one document is in the source language of the user query and the other in the target language. Moreover, the document pairs are translations of each other. In some approaches the documents are not exact translations of each other but nevertheless about the same topic (comparable document collections).

Figure 1. Parallel corpora in CLIR

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 then their target language pairs are 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, which then is run against the target language collection.

The corpus-based approach requires a parallel collection in the domain of the queries and the target collection. It can be much smaller than the target collection but needs to contain pertinent vocabulary for the query topics. There is no word-by-word correspondence between the source and target language queries. In fact, the latter contains words that are statistically associated with the source language query words in the parallel corpus.

Because electronic parallel corpora are not readily available in the different languages, this CLIR approach is not a practical option in the South African context.

2.2 Machine translation

In this approach, a machine translation system is employed for query translation. Such systems aim at correct target language translation of source language texts. Translation is based on translation dictionaries, other linguistic resources and syntax analysis to arrive at an unambiguous and high-quality target language text (Figure 2). The source language query in CLIR applications must be a grammatically correct sentence (or a longer text) for the translation to be successful.

A major problem for this approach is the availability of good machine translation systems. For many language pairs no systems exist, and for many others, the quality of the systems is rather poor and/or their topical scope limited. In the

South African context, there is a machine translation system between English, Afrikaans and several African languages (*http://lexica.epiuse.co.za*) but its quality for CLIR applications is inadequate.

Figure 2. CLIR based on machine translation

2.3. Dictionary-based techniques

When the linguistic resources for a language pair are limited, dictionaries are the most likely to be available in machine-readable form. Therefore one may say that the dictionary-based approach is viable when linguistic resources are scarce.

The dictionary-based approaches use bilingual dictionaries for query translation (Figure 3). A very basic strategy for query translation is to process it word-by-word and, for each source language word, look up its target language equivalents and put them into the target language query.

Figure 3. Dictionary-based CLIR

No translation dictionary available for CLIR is as extensive as the lexicon of the language listing all its words. This is because natural languages are productive systems that continuously generate new words. The most important categories of untranslatable query keys generally not found in general dictionaries are new compound words, proper names and other spelling variants, and special terms (Pirkola *et al.*, 2001). Pirkola *et al*. (2002) show how n-grams can be used to match the untranslatable source language words with similar words in the target database index. We will consider this in more detail in the next subsection.

If the source language words appear in inflected forms, they cannot be readily translated, because they do not match with dictionary headwords (which are in base forms). If there is a morphological analyser (parser) available, words can be normalised to the lemma, i.e. the normal dictionary entry form. In the case of the South African indigenous languages no such parsers are available and therefore the index has to be in inflected word form. Promising work on such parsers is in progress for some languages at the Universities of Stellenbosch and South Africa (http://www.ast.sun.ac.za). The

possibility that a full set of parsers will be available for all languages in South Africa in the near future is, however, slight. Therefore, the possibility of having indexes available in normalised form, based on morphological parsers, is not likely. In the South African context the retrieval techniques therefore have to work on an index containing words in their inflected form from multiple target languages and matching can be achieved through approximate string matching.

2.4. Approximate string matching in IR

A common method to handle untranslatable words, e.g. many proper names, in dictionary-based CLIR, is to pass them as such to the target language query. However, in the case of spelling variants or inflectional forms, a source language form does not match the variant form in the target database index, causing loss of retrieval effectiveness. To find target language spelling variants for untranslatable source language words, some *approximate string matching* technique, such as *n-gram based matching*, *edit distance* or *LCS* (longest common subsequence) technique has to be applied.

In *n-gram matching* query keys and index entries are decomposed into n-grams, i.e., into the sub-strings of length n, which usually consist of the adjacent characters of strings. *Digrams* contain two and *trigrams* three characters. The degree of similarity between query keys and index entries can then be computed by comparing their n-gram sets (Pfeifer *et al*., 1996, Robertson & Willett, 1998, Salton, 1989, Zobel & Dart, 1995).

Pirkola *et al*. (2002) showed that n-gram matching is effective in searching for cross-lingual spelling variants. It can also be used an alternative method for stemming algorithms (Xu & Croft, 1998).

N-gram matching is a *non-metric* approximate matching technique. A *metric* similarity measure *distance* is a similarity function, which satisfies the following conditions (Berghel, 1987):

1. *distance*(s, s') >= 0
2. *distance*(s, s') = 0 <--> s=s'
3. *distance*(s, s') = *distance*(s', s)
4. *distance*(s, s') + *distance*(s', s'') >= *distance*(s, s'')

for arbitrary character strings s, s', s''.

Metric similarity measures involve *edit distance* and *LCS*. Edit distance is defined as the minimum cost required to convert one string into another. Conversion includes character insertions, deletions and substitutions. For example, the minimum number of steps required to change the string *industry* into *interest* is six (Kruskal, 1983). For two words, their LCS is the longest character sequence of the sequences that occur in both words. For example, for the words

r e t r i e v a l and *r e v i v a l* LCS is *r e i v a l*.

When using approximate matching in IR, the query key is matched against database index entries and the best matching entries can be added to the query. These will then be treated as a synonym list. The number of entries added may be limited, either through a threshold value (the calculated value of the similarity must exceed a set limit value for the index entry to be accepted), or by adding only the *k* best-matching index entries to the query, or by both criteria.

3. The viability of using CLIR to access Zulu language databases

3.1.The Zulu language

Zulu is spoken by more than 8,8 million people in South and Southern Africa. Of all the languages (including Afrikaans and English) spoken in South Africa, Zulu has the largest number of speakers. Zulu is an agglutinative language, which means that grammatical information is conveyed by attaching prefixes and suffixes to roots and stems. All Bantu languages are divided into classes or sets, called grammatical genders. Each gender has two distinct prefixes, one for singular and one for plural nouns. The classes far exceed the familiar European classifications of masculine, feminine and neuter, and are roughly associated with semantic characteristics relating to, for instance, human beings, kinship terms, animals, plants, artifacts, abstract concepts and so on. *Umlimi* and *abalimi* are, for example, the singular and plural forms of the noun meaning *farmer*, and most words denoting human beings are in the *umu-*, *aba-* class (Class 1) of nouns. Verbal nouns, on the other hand, belong to the Class 8 group of nouns, of which the prefix is *uku-*, for example *ukugula*, which is the Zulu word for either *death* or the infinitive *to die*.

Verbs are complex - a system of affixes mark the different grammatical relations, such as subject, object, tense, aspect and mood. For example, suffixes on verbs are used to derive passive, active, causative, reciprocal and prepositional verb forms. There is a system of concordial agreement in which subject nouns, object nouns and other words must agree with the verb of the sentence in class and number. Adjectives, possessive nouns and demonstratives also agree with the noun that they modify in class and number.

The phonology of Zulu is characterized by a simple noun inventory and a highly marked consonantal system with ejectives, implosives and click-sounds. It also is a tone language with inherent low and high tones (Doke, 1990, "UCLA", 2002).

3.2. Methodology

In order for an English-speaking person (source language) to access an IK database in Zulu (target language), the process may be described as follows (see Figure 4):

Figure 4. English to Zulu CLIR Process

First the English query is translated into the target language, Zulu (one pair). (This applies for all pairs necessary.) The translation can be done by using the dictionary translation method. Since there are many morphological analyzers for English available, it is trivial to match the English words in natural language to the translation dictionary entries. For each English word we then get a number of Zulu translations, some of which are correct and some of which are incorrect for the query context (due to the ambiguity of natural language). All these words are then matched against the inverted index of the Zulu database. The inverted index is not normalized because there are no suitable Zulu morphological analyzers available. Therefore, approximate matching between the query words and indexed words is necessary. For each Zulu word (base form), a number of best matches in the index are identified. These are treated as synonym sets and query structure (Pirkola *et al.*, 2001) is utilized to reduce ambiguity.

Some parts of this process have been tested empirically on a corpus, and this will be described below, including some of the problems experienced so far. Due to the fact that there are no large-scale Zulu language databases available, the reverse of the process is being tested, that is, Zulu queries put to an English language database (Figure 5).

The corpus used was the CLEF English document collection with 50 topics (CLEF 2001). The title and description fields were translated from English into Zulu by independent translators. As there were no Zulu-English bilingual translation dictionaries available in electronic format, part of a printed dictionary had to be retyped manually in a word-processing program. Due to various restraints and restrictions, it was decided to only create a monolingual word list (Zulu entries only) in electronic format. The dictionary used was the 1990 edition of the Zulu dictionary by Doke *et al.*, where singular forms, plural forms as well as stems for nouns are listed. Using this particular dictionary should thus alleviate ambiguity problems for nouns. Verbs, however, are listed only as stems.

We tested the following approximate matching techniques to match the individual Zulu words in the topics with Zulu dictionary entries: (1) conventional digrams and (2) conventional trigrams, i.e., digrams and trigrams combined of adjacent letters of words, (3) classified s-grams, (4) edit distance, and (5) LCS.

The classified s-gram matching technique (where *s* refers to the term *skip*) refers to a novel n-gram matching technique described in detail in Pirkola *et al*. (2002). In the technique, digrams are combined both of adjacent and non-adjacent characters of words. Digrams are classified into categories on the basis of the number of skipped characters. Digrams belonging to the same category are compared with one another, but not with digrams belonging to a different category.

Figure 5. The reverse process tested

In the case of n-gams (the cases 1-3 above) similarity values were computed using the following string similarity scheme (Pfeifer *et al*., 1996):

$$SIM(N_1, N_2) = |N_1 \cap N_2| / |N_1 \cup N_2|,$$

where N1 and N2 are n-gram sets of two words. ¾N1 « N2¾ denotes the number of intersecting (similar) n-grams, and ¾N1» N2¾ the number of unique n-grams in the union of N1 and N2. For example, the degree of similarity for the words rwanda and ruanda is calculated as follows (for conventional digrams):

SIM({rw,wa,an,nd,da},{ru,ua,an,nd,da})

= |{an,nd,da}| / |{rw,wa,an,nd,da,ru,ua}| = 3 / 7 = 0.429.

Table 1. Matching results for 65 Zulu source words in CLEF Topics C041 to C045

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	

Five of the CLEF 2001 topics were used as a trial to establish which of the procedures described above would give the best results. For topics C041 to C045 there were 75 Zulu source words. Ten of these were proper nouns, which were

not matched, and therefore not used in the calculations. The results were thus based on 65 source words. For each procedure the six approximate best matches were listed for each of these source words. It was then manually established which one of these six words was the correct match for the source word. If the first word was the correct match, a value of 1 was allocated, a value of 2 if the second word was the correct match, and so on, and if there was no match, a value of 7 was given. The results are given in Table 1.

Translations from the Zulu **source** words into English were again done manually, following strict mechanical rules (all senses collected). The first type of problem experienced in this phase related to long English descriptions needed to paraphrase single Zulu words, and these had to be marked-up as phrases. Examples of these are the Zulu nouns *idlingozi* which is paraphrased as *an outburst of intense interest* and *isinyabulala* which is paraphrased as *a person weak from age.* This problem is also a result of the disparate vocabularies between "Western" and indigenous languages described below. The second problem was related to homonyms, where one word has various meanings, for example the **source** noun entry - *zwe isizwe izizwe* can mean either *tribe* or *a rapidly spreading brain disease*.

The retrieval results of the translated Zulu queries in the English collection are reported in a later paper.

3.3. Problems regarding resources

3.3.1.No electronic dictionary

At present there are no (comprehensive) electronic dictionaries for any of the African languages of South Africa. Various projects of the different dictionary units are in the process of being established; however, nothing that can be used for research is available at this stage. We therefore used a retyped monolingual dictionary with manual translations from Zulu to English, as explained above.

3.3.2.Test database - CLEF suitability

The aim of the research is to test whether English search strings could be run against a Zulu database, and to see which technologies are possibly viable in this regard. Since no Zulu database was available, we had to reverse the process, i.e. to run Zulu queries, translated into English, against an English database. We used the CLEF 2001 database, as explained above. We therefore tested whether inflected Zulu words can be matched with a dictionary entry through approximate matching. This proved to be fairly successful, as indicated in table 1. The reverse process, i.e. to match dictionary entries to inverted indexes that contain non-normalized, inflected text, should therefore also be possible. However, the process indicated another interesting problem that should be investigated in its own right, *viz.* the impact of the lack of technical terminology in Zulu CLIR. The CLEF database and queries are very Euro-centric, and many of the search terms could not be translated directly into Zulu.

The problems in such matching are classified and discussed below.

3.4Problems experienced with matching

Zulu does not have single word translation equivalents for many technical and scientific terms. This problem is solved either by paraphrasing to explain concepts or by borrowing words from other languages.

3.4.1. Paraphrasing

Paraphrasing is very commonly used to describe words or phrases in technical or scientific terminology. Two examples will suffice:

> CLEF 2001 topic C041 contains the word *pesticides*. The Zulu translation of this reads *amakhemikheli abulala zonke izifo enzinengozi* which translated directly back to English means *chemicals that kill all illnesses which are dangerous*.
>
> CLEF 2001 C046 topic contains the phrase *embargo*, which is translated into Zulu as *ukuvimbela kohwebo*, which literally means *to prevent trade.*

Mechanically, n-grams match most of the individual words quite well to the dictionary entries. On a conceptual level however, the result is not always very good.

3.4.2. Borrowed words

Zulu has borrowed extensively from Khoisan (the languages of the southern African aboriginal hunter-gatherer populations) and in modern times from English and Afrikaans. These borrowed words may take one of two forms - the word is either kept as in the original language, but prefixed as if it is a Zulu noun, or assimilated into the Zulu language. Examples of the latter are the English nouns *bicycle* which becomes *ibhayisekili* in Zulu, *post office* becomes *iposihovisi* and *tea* becomes *itiye*. Even though the spelling is vastly different from the English, the pronunciation is similar to the original. Some of these borrowed words are of acceptable use and are included in the dictionaries. However, some are new additions, and the modern concepts are very often not entered in dictionaries (which are not updated very often). An

example in the CLEF topic set is *amakhemikheli* (consisting of the prefix *ama-* indicating that it is an artifact and -*khemikheli* which is borrowed from the English word *chemicals*) which was not found in the dictionary.

Some of the words are not assimilated into the language, but stay in the original language from which it was borrowed, with a Zulu class-prefix added. Examples which occurred in the translated CLEF topic set are, *inter alia, ama"computer viruses"*, *ama"computer mouse"*, *i"film festival"*, *i"green power"*, *i"mad cow disease"*, *i"German property speculator"* and *iTurqoise*.

In the first six examples above, the prefix is not added to a single word, but to an entire phrase and the whole phrase is put in inverted commas, which may compound the problem of identifying phrases in CLIR.

3.4.3.Inflected word forms

If the source language words appear in inflected forms, they cannot be readily matched, because they do not match the dictionary base forms. As stemming is not an option in Zulu due to the lack of morphological parsers, approximate matches have to be made by n-gramming or similar string-matching techniques.

Below are two examples of approximate matching – the first of a noun and the second of a verb:

The Zulu monolingual dictionary entries for nouns are of the format:

stem, singular (prefix + stem), plural (prefix + stem)

The entry for *danger* thus is:

-ngozi ingozi izingozi

In one of the translated CLEF topics the inflected form is *ezinengozi.* This inflected form was matched as follows by the five different matching techniques:

Trigram: n=3, LCS: n=1, Edit: n=7 (no match), Skipgram: n=4 and Digram: n=2

In the case of verbs the stem only is listed, e.g, *bulala* for *to kill*. In the running text of a topic it appears as *abulala* (in concord with the noun *amakhemikheli*). This was matched as follows:

Trigram: n=2, LCS: n=2, Edit: n=2=3, Skipgram: n=2 and Digram: n=3

Although these two examples were arbitrarily chosen, they do show that inflected word forms have a significant influence on accuracy of retrieval, due to the increased possibility of noise.

Even though the CLIR techniques discussed above provide significant results there are a number of problems due to *inter alia* noise, untranslatability, paraphrasing, conceptual mismatching, problems inherent to the Zulu language, etc. The refinement of CLIR techniques may result in improved recall and precision in future. However, in practice, such as IK databases, precision could be insufficient, and we would suggest that further techniques be employed to improve it. A practical solution would be to employ metadata. It is therefore also necessary to look at metadata as a mechanism to improve precision in conjunction with CLIR.

4. Metadata as filtering mechanism

The content of a knowledge source may be described though metadata. The purpose of metadata is thus to describe the structure of the data, and more importantly, to capture any additional properties that may characterize it.

The *de facto* standard for metadata, especially on the Web, is Dublin Core. The elements and definitions of Dublin Core as listed in table 2, columns 1 and 2 are based on the official standard for the element set of DC (ANSI/NISO Z39.85-2001). The elements can be seen as describing three different dimensions of metadata, viz describing the content or ***data***, describing the ***source***, and describing the ***collection process*** to collect the content, as given in column 4 of table 2. This subdivision is very relevant for IK - and may be relevant in other cases as well - since it describes the aboutness, isness and processing of the information objects.

Dublin Core is a general set of metadata elements and is very often enriched by application domain dependent additions, such as the NDLTD (Networked Digital Library of Theses and Dissertations http://www.ndltd.org) and the LOM (Learning Object Metadata http://ltsc.ieee.org/wg12/). It would therefore be natural to extend the metadata set for the description of IK objects, and we propose the additions as listed in column 3 of Table 2. The detail of these additions are described elsewhere (Cosijn *et al.*, 2002). These additions are the minimum that are required, and more can be proposed.

The proposed metadata structure provides a multi-viewpoint access to the content. However, it is limited and constrained by all the limitations of bibliographic databases and controlled vocabulary access, such as incomplete indexing, incorrect and inconsistent keywords, uninformative abstracts, missing data, lack of thesauri and controlled vocabulary lists, etc.

The needs that the users may have are unpredictable and therefore one cannot outline any finite element set that would cover the needs. It might be possible to develop and populate such a set, but this becomes economically unfeasible. This is a complex area, and needs further investigation.

We may use the following example to illustrate the value of using both metadata and full-text access to increase precision in retrieval of IK: a user is looking for information on herbs, specifically on the use of herbs for curing fever. By specifying, for instance, the proposed DC field subject.domain in the above example as *medicine,* items dealing with *agriculture* will not be retrieved. The effect of this is that results will be limited to the domain of medicine only, and therefore potentially ambiguous results outside the field of medicine will not be included in the search results.

Table 2. Description of DC with proposed added subdivisions

Name of element	Definition	Proposed added subdivisions	Dimension of description
Title	Name given to the resource		Data
Creator	Entity primarily responsible for making the content of the resource	creator.status	Source
Subject	Topic of the content of the resource	subject.keyword subject.domain	Data
Description	Account of the content of the resource		Data
Publisher	Entity responsible for making the resource available		Collection process
Contributor	Entity responsible for making contributions to the content of the resource	contributor.recorder contributor.editor	Collection process
Date	Date associated with an event in the lifecycle of the resource	date.recorded date.published date.added	Collection process
Type	Nature or genre of the content of the resource		Source
Format	Physical or digital manifestation of the resource		Source
Identifier	Unambiguous reference to the resource within a given context		Source
Source	Reference to a resource from which the present resource is derived		Source
Language	Language of the intellectual content of the resource		Source
Relation	Reference to a related resource		Source
Coverage	Extent or scope of the content of the resource	coverage.location coverage.timespan coverage.tribe	Source / Data
Rights	Information about rights held in and over the resource		Source
Proposed addition		funding.agency	Collection process

Source: Based on http://dublincore.org.

5. Conclusion

The contributions of the present paper are:

- An outline for an approach for information retrieval in IK databases based on CLIR and metadata techniques.
- An outline of the CLIR process, involving approximate string matching, when the target (or the source) language is an indigenous language lacking linguistic resources and exhibiting strong inflectional morphology.
- An analysis of approximate string matching success in identifying inflected Zulu word forms based on their base forms (dictionary forms).
- An analysis of the translation problems between an indigenous language and a western language.
- An outline for the application of the Dublin Core metadata format for IK databases.

We have shown that there are many limitations inherent to querying the full text of IK databases if we rely solely on CLIR techniques, such as ambiguity, incorrect stemming, paraphrasing in translations, untranslatability, conceptual mismatching, etc., which may have a negative impact on the quality of retrieval. On the other hand, there are a variety of limitations impacting on the quality of search results when querying an IK database using metadata only, for example, low precision due to broad keywords, lack of synonyms in the query, etc. It may, however, be possible to improve both precision and recall, as required, by querying content in combination with metadata searching, and we therefore recommend that a bilateral approach is followed to enhance accessibility of IK databases.

Regarding CLIR techniques, we outlined the process translating queries from English to Zulu. The translation of English words into Zulu base form words is for a large part manageable while it presents sometimes difficult problems of conceptual incompatibility between the languages. However, matching the Zulu base forms to the Zulu inflected form index, due to the lack of morphological analyzers, was a novel challenge. We outlined a process based on monolingual approximate string matching in Zulu to identify the inflected forms of query words in the database index. However, it was not possible to test the process directly because there are no large-scale Zulu databases available. We therefore tested the opposite, and focused on the translation problems of Zulu topic words into English queries. The first step of identifying correct base forms for Zulu was already problematic. While approximate string matching gave relatively good results, they were far from perfect (not all words were translated, the correct base forms were not always top-ranked, etc). This introduced ambiguity into the translation process. Typical matching problems at this stage were unmatched proper nouns, inflected word forms, borrowed words and paraphrased translations.

Translation of the identified Zulu base forms (several for each topic word) was done manually but simulating a mechanical process based on a translation dictionary. In this process, a number of problems between the languages turned up. Problems encountered at this stage were mainly related to the paraphrasing of the Zulu to English translations, and the translations had to be marked-up as English phrases. Similar problems are encountered in the opposite direction of translation.

Present research in CLIR concentrates on languages with comparable vocabularies in terms of, for instance, technical and scientific terminology. This research has shown that a set of new problems will be encountered if the language pairs used contain disparate vocabularies and this increases the complexity of CLIR. The complexity is increased when one of the languages dealt with lacks resources for word form analysis. These problems need further investigation, and techniques will have to be found to deal with this. This may present unique opportunities for research in CLIR.

Declaration

This paper was presented at the CoLIS 4 Conference (Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods 2002), 21 - 25 July 2002, Seattle WA, USA

References

Berghel, H.L. (1987). A logical framework for the correction of spelling errors in electronic documents. *Information Processing and Management, 23(5),* 477-494.

Cosijn, E., Järvelin, K., Bothma, T., Nel, J.G. & Theophanous, J. (2002). Facilitating access to knowledge databases in indigenous languages. In: *Proceedings of the 15th Standing Conference of Eastern, Central and Southern African Library and Information Associations. SCECSAL 2002* (to appear).

Doke, C.M., Malcolm, D.M., Sikakana, J.M.A. & Vilakazi, B.W. (comp) (1990). *English - Zulu Zulu - English Dictionary*. Witwatersrand University Press, Johannesburg.

Hull, D. & Grefenstette, G. (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland*, pp. 49-57.

Kruskal, J.B. (1983). An overview of sequence comparison. In: Sankoff, D & Kruskal, J.B. (Eds.) *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley, Reading, Massachusetts.

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

Pfeifer, U., Poersch, T. & Fuhr, N. (1996) Retrieval effectiveness of proper name search methods. *Information Processing & Management, 32(6):* 667-679.

Pirkola, A., Hedlund, T., Keskustalo, H. & Järvelin, K. (2001) Dictionary-based cross-language information retrieval: problems, methods and research findings. *Information Retrieval 4 (3/4)*, 209-230.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, H. & Järvelin, K. (2002) Targeted s-gram matching: a novel n-gram matching technique for cross-and monolingual word form variants. *Information Research* (to appear).

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

Salton, G. (1989) *Automatic text processing: the transformation analysis and retrieval of information by computer*. Addison-Wesley.

UCLA Language Materials Project. (2002) *Zulu Profile*. Retrieved January, 9, 2002 from: http://www.lmp.ucla.edu/profiles/profz01.htm

Xu, J. & Croft W.B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems, 16(1),* 61-81.

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