Bothma, T.J.D. & Kaniki, A. 2004. *ProLISSA 2004. Proceedings of the 3rd biennial DISSAnet Conference, Pretoria, 28-29 October 2004.* Pretoria: Infuse, pp. 97-110.

Afrikaans-English cross-language information retrieval

ERICA COSIJN[1], HEIKKI KESKUSTALO[2], ARI PIRKOLA[2], KAREN DE WET[1] & KALERVO JÄRVELIN[2]

[1]Department of Information Science, University of Pretoria, Pretoria, South Africa &
[2]Department of Information Studies, University of Tampere, Tampere, Finland
Email: erica.cosijn@up.ac.za, ccheke@uta.fi, ari.pirkola@uta.fi, karen.dewet@up.ac.za & kalervo.jarvelin@uta.fi

This study reports on the first experiments ever to apply dictionary-based query translation techniques to Afrikaans queries submitted to an English database. The system was evaluated using 35 topics from the CLEF 2001 English language collection (title and descriptions). To show the performance level of the test queries, the original English queries were run as baseline queries. They contained the title and description field words of the CLEF topics as query keys, and the test queries were formed on the basis of the same words. Combining a bilingual dictionary, a morphological analyser and a stopword list, the Afrikaans queries were translated into English and matched with the English target database. The engineering of the translation dictionary and the creation of a stopword list are described. Morphologically, the nature of Afrikaans is quite simple and therefore an Afrikaans morphological normaliser for information retrieval was developed as part of this study. The results of the test runs and an analysis of errors encountered are also reported.

1. Introduction

The basic idea of cross-language information retrieval (CLIR) is to bridge the language boundary by providing access in one language (the source language) to documents written in another language (the target language), by using query translation from the source language into the target language and/or document translation from the target language. The main strategies for query translation are based on three different methods, namely dictionary-based methods with specific relevance to (bilingual) translation dictionaries, corpus-based methods, and machine translation (Oard & Diekema, 1998; Pirkola et al., 2001).

This study utilises query translation by means of a bilingual translation dictionary to translate Afrikaans (the source language) queries into English (the target language), and then matches the latter to an English language database, with English language documents as the retrieved results. The process is summarised in Figure 1.

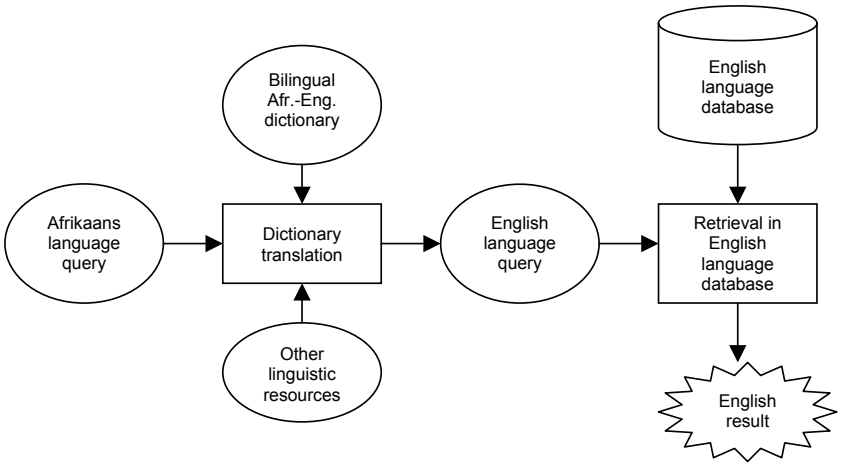

Figure 1: Query translation from Afrikaans to English using a bilingual dictionary

The Afrikaans-to-English query translation is based on the UTACLIR (University of Tampere Cross-Language Information Retrieval) framework described in Hedlund et al. (2004). The main features adopted in this study include:

1. Normalising the target index
2. Utilising source and target language stopword lists
3. Normalising the source keys
4. Splitting the source compounds and recognising their components when they cannot be translated as a whole
5. Using a bilingual translation dictionary
6. Normalising the target keys;
7. Performing approximate string matching for untranslatable source keys
8. Structuring the final target query

This paper investigates the following basic research question: Based on the dictionary translation approach, how can Afrikaans information requests be translated into English queries, and how effective are such queries in retrieving documents in an English collection?

There are no bilingual translation dictionaries or morphological analysers designed for CLIR in Afrikaans, as well as no stopword list. Therefore, the sub-problems that had to be addressed in this research were:

- How should Afrikaans words be normalised for matching with the dictionary headwords?
- How should the translation dictionary be organised to facilitate CLIR?
- What is a suitable stopword list?
- What are the components and order of execution of the translation process?

- How effective are the processes of normalisation and translation, and what are the pitfalls and problems?

This paper describes the first Afrikaans-to-English CLIR system developed utilising query translation through a bilingual dictionary, and reports on the effectiveness of the morphological analyser developed specifically for this project. The results are analysed and the errors categorised in order to make recommendations for future system development.

2. Methodology

2.1 Morphological structure of the Afrikaans language

Afrikaans is one of 11 official South African languages. It developed mainly from Dutch and is a Germanic language. In order to create a normaliser to match the Afrikaans word in running text to the dictionary headwords, the morphological structure of the Afrikaans language had to be established. This was done by using a corpus of Afrikaans newspaper articles which contained roughly 3 500 words. The process was as follows:

1. Each of the unique words (1 072) in the corpus was checked against the dictionary and base forms, and stopwords were identified.
2. Plural forms appearing in the text were listed and analysed.
3. Compounds appearing in the text were identified and checked against the dictionary. If these compounds did not appear as headwords, they were matched against dictionary headwords to recognise the component words of compounds.
4. Past tense forms and other inflected words were listed and categorised.
5. Proper names appearing in the text were listed separately.
6. All other words not listed above were analysed and classified.

The analyses confirmed that the Afrikaans language is morphologically rather simple. The results are categorised in Table 1.

The implication of this analysis is that approximately 88% of words in Afrikaans running text (entries in the table marked with an asterisk in the last column) should be found in the dictionary either as is, or by applying a simple morphological analyser to normalise the word form (see Section 2.4 below).

2.2 Dictionary filtering

A commercial bilingual Afrikaans-English dictionary was used. In order to adapt it for CLIR purposes, the following processes were applied:

1. The headwords were identified by string-based rules.
2. The alternative spellings were identified and listed as separate headwords, e.g. *oorkrabbertjie:* earring*; oorkrawwertjie:* earring.

Table 1: Manual analyses of Afrikaans text corpus

Category	n	%
Stopwords	150	14,0%*
Headwords	565	52,7%*
The plurals of words containing a double vowel *(oo, aa, ee, uu)* may be formed by dropping one of the vowels and adding a suffix *-e* to the end of the word (e.g. *boom* > *bome)*	18	1,7%*
Plurals formed by adding a suffix*: -e, -s* or *-'s*	85	7,9%*
Other suffixes attached to the headword (e.g. *-es, -de, -ste, -ie, -pie,* etc.)	59	5,5%*
Past tense forms containing the prefix *ge-*	13	1,2%*
Compounds (with or without the fogemorphemes *-e-* or *-s-)*	58	5,4%*
Past tense form *-ge-* embedded in the word	26	2,4%
Other (e.g. misspelt in the original text)	18	1,7%
Proper nouns	80	7,5%
Total:	1 072	100,0%

3. In cases of homonyms, word senses were separated and listed individually under headwords, each sense of the word having its own headword, e.g. *saal:* hall; *saal:* ward; *saal:* saddle.
4. Headwords with options for the formation of compound words were automatically identified, then merged into their various possible compounds. Each compound was listed as a new headword with translations, e.g. *meubel:* piece of furniture; *meubelfabriek:* furniture factory.
5. Plurals were not included in headword formation in these initial runs, but were identified by a simple morphological analyser (Section 2.4).
6. As a final step, the dictionary was checked manually and fine-tuned.

At the end of this process, the dictionary contained more than 160 000 unique Afrikaans words and phrases, with their English translations attached.

2.3 Stopword list

A stopword list containing 341 words was created by translating an existing English stopword list into Afrikaans, as well as calculating the word frequency in a corpus of Afrikaans newspaper articles. These two resources were compared and merged to create a final stopword list. Thereafter, the list had to be manually checked and modified, due to a relatively high incidence of homonyms occurring either as stopwords or as content-bearing words, e.g. *deur* means either "through" (a stopword) or "door" (content-bearing); or *van* means either "of" (a stopword) or "surname" (content-bearing).

A further complication in creating a stopword list was the unique characteristic that Afrikaans vowels in basically any word may be accented for emphasis, and that the accented word is not an entry in the dictionary, e.g. *daar* and *dáár* both mean "there",

but the latter is emphasised to indicate focus to the reader. All stopwords that could conceivably be emphasised in running text were therefore duplicated in the stopword list, both as an accented and as a non-accented word. (See Section 5.6 for a discussion on incidental diacritics in Afrikaans.)

2.4 Normaliser

Our goal was to translate the source queries by utilising the translation dictionary. The user-given keys may be expressed in plural or prefixed forms, while the translation dictionary typically contains words only in their basic forms. Moreover, the user-given compounds may be missing from the dictionary. Therefore, we needed to be able to normalise the source keys and split the compounds in order to find dictionary matches. To solve these problems, a simple Afrikaans normaliser was designed.

Moreover, the source queries may contain non-content-bearing words, such as prepositions, which must not be translated. Finally, the source queries may contain untranslatable expressions, like proper names. To solve these two problems, we utilised stopword lists for non-content-bearing source and target keys (Section 2.3), and approximate string matching for untranslatable source keys (see Hedlund et al., 2004).

The normaliser depicted in Figure 2 utilises a word list containing approximately 82 000 unique Afrikaans single-word entries. The ISO 8859-3 character set is used for expressing the characters with diacritics. The normaliser automatically categorises the input string belonging to exactly one of seven distinct key types. This process is described next.

The input key is first compared with the word list in order to find an exact match. If a match is found, the key is recognised and returned by the normaliser (Key Case 1). For example, the key *daarna* is found in the word list and thus will be returned by the normaliser.

Generally speaking, if no direct match is made, the string is modified by several successive steps, e.g. a plural suffix is removed and further matches are attempted by using the modified strings. The process is continued until a match is made, or the key is classified as "unrecognised". For example, the uppercase form *Vrees* does not belong to the word list, thus it is transformed into a lowercase form, after which a match is made (Case 2) and the matched form *vrees* is returned. In Case 3 (*ge-* prefix), for example, the word *gedoen* does not exist in the word list, but after removal of the prefix a match is made for the remaining string, and *doen* is returned. In Case 4, *bome* does not exist in the word list, but after removing the ending *-e* and doubling the remaining last vowel *o*, a match is made to *boom*.

Starting from the left-hand side of the source key, compounds are split by finding the longest common subsequence (LCS) from the word list. Fogemorphemes (joining morphemes) *-e-* and *-s-* are also taken into account. For example, the key *kliëntekontak* is not found in the normaliser word list. The longest subsequence found is *kliënt*. Failing to match *ekontak,* we next attempt a match without the fogemorpheme *-e-* and find a

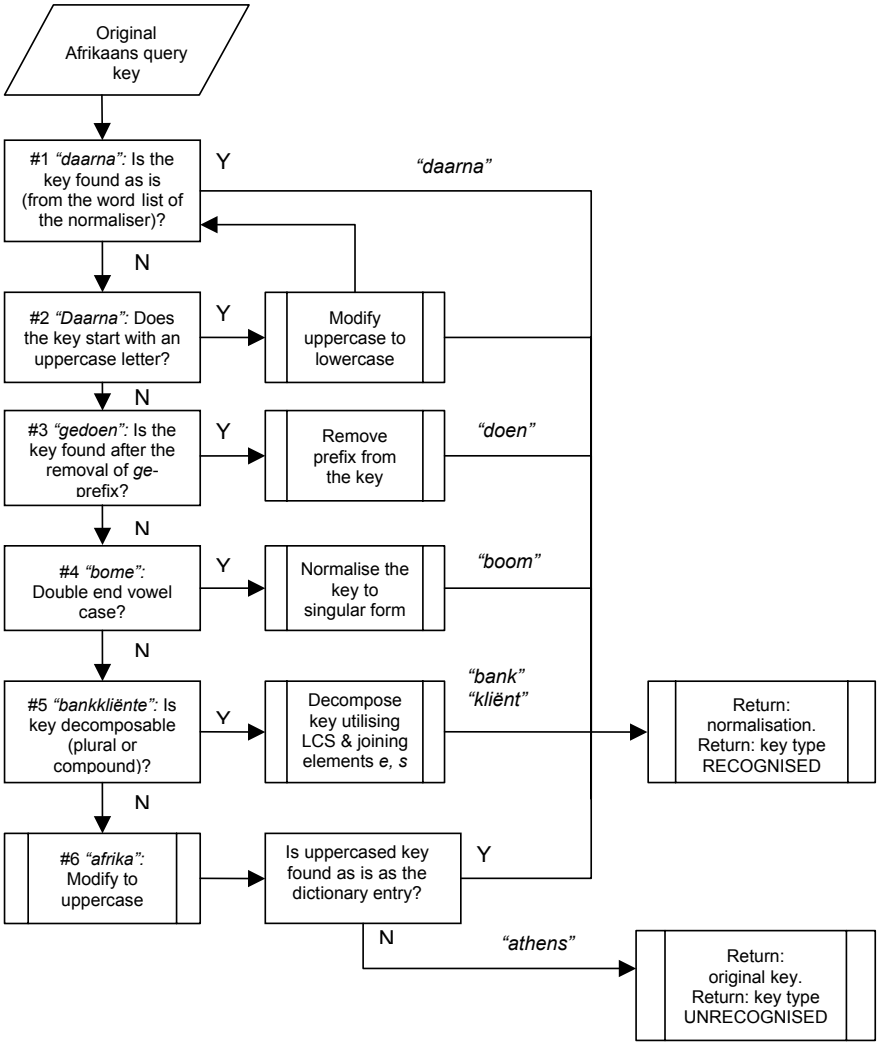

Figure 2: Simplified Afrikaans normaliser (sample words in quotes)

match with *kontak*. The component words are then returned by the normaliser (Case 5). As the last "recognised" case, it is possible that a user-given lowercase word exists in the word list but only in capitalised form. For example, the input key *afrika* is not found, and as the last step of the normalisation process, we try to match its uppercase form *Afrika*, this time successfully (Case 6).

The final case of the normalisation, Case 7, applies to keys not matched by any of the described attempts, e.g. the string *athens* cannot be recognised even by suffix stripping or LCS matching. In Case 7, the status "unrecognised" is returned by the normaliser, as the query translation process needs this information to apply fuzzy matching for unrecognised keys.

2.5 Query translation

The main components of the Afrikaans-English UTACLIR system are:

- Source and target language normalisers
- A translation dictionary
- Source and target language stopword lists
- An approximate string matching system

The source normaliser was described under Section 2.4. As a target normaliser, ENGTWOL by Lingsoft was used for normalising the dictionary translations. The translation dictionary contains approximately 160 000 Afrikaans entries, including repetitions. A source stopword list (341 Afrikaans words) was developed to avoid translation of non-content-bearing source keys (Section 2.3). The source normalisation process guarantees that the possible uppercased stop keys also are omitted, although the stop list contains only lowercase keys. A target stop list (758 English words) was also included in the system, as the dictionary translations themselves may contain non-content-bearing words (see below). An approximate string matching module developed in an earlier study was included as part of the system (Hedlund et al., 2004).

The translation process (Figure 3) includes two alternative main branches: one for Afrikaans keys that are recognisable by the source normaliser (Cases 1-6 described in Section 2.4), and another one for the unrecognisable keys (Case 7).

Recognised source keys are also translatable, because the translation dictionary includes all words in the word list of the normaliser. Three cases are recognised:

- First, if the key is a source stopword, we simply remove it (or actually express it in the target query by a non-matching expression *#syn(nullstr)*).
- Secondly, if the key is not a stopword and the normaliser returned one word form, we translate it, normalise the translations and remove the target stopwords from the translation. For example, the dictionary translation of *voel* includes the translation "be aware of". The words "be" and "of" are removed from the target query, as they belong to the target stop list.
- Third, if the key was not a stopword and the source normaliser returned two or more word forms (due to compound splitting), we translate all components, normalise the translations, and finally remove the target stopwords. For example, for the source key *kliëntekontak* the normaliser returns *kliënt* and *kontak*, which are translated as "client" and "contact".

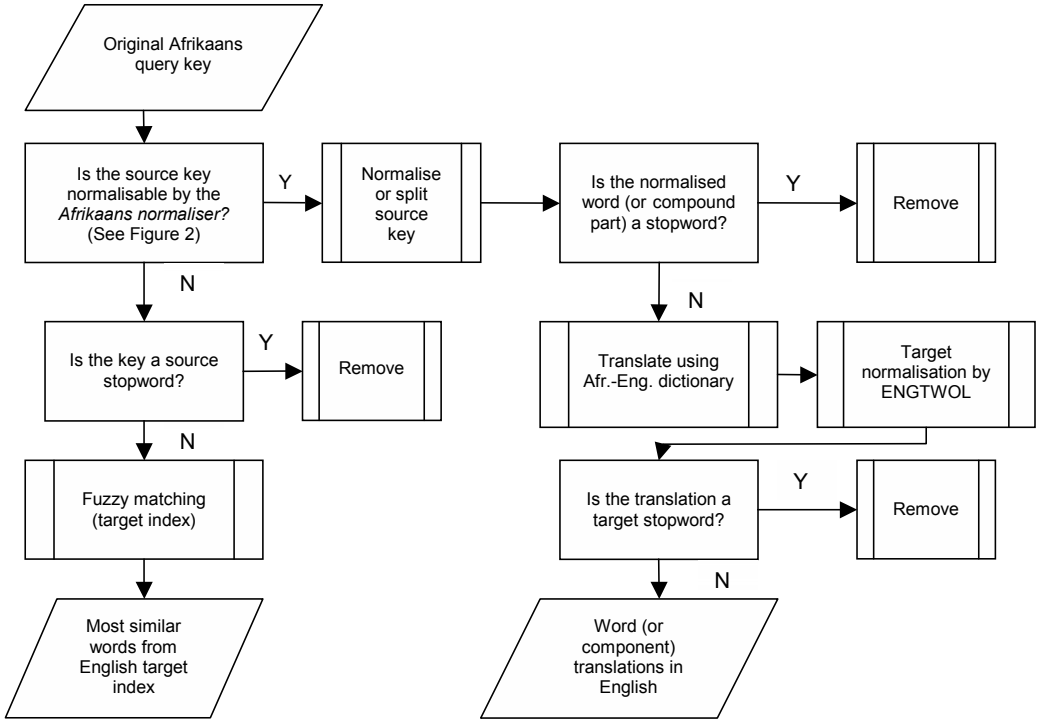

Figure 3: Afrikaans-English query translation. The translation process includes source and target normalisation, source and target stopword removal, and dictionary translation or fuzzy matching.

Simple query structuring is performed for the translations so that one synonym *(#syn*) statement in the target query corresponds to an individual source key (see the explanation of these operators under Sections 3.1 and 3.3). For example, for the source key *kliëntekontak,* a target statement *#syn(client contact)* will be formed.

In the case of an unrecognised key, the key returned by the normaliser does not exist in the word list of the normaliser or the translation dictionary, thus it is also untranslatable. If the unrecognised source key is a stopword, we remove it. If the key is not a stopword, we use approximate string matching to select the two best matches from the target index.

For example, for the unrecognised (and untranslatable) Afrikaans key *MacDonalds,* the words *macdonald* and *@mcdonalds* are retrieved from the target index and added to the final target query *(#syn(macdonald @mcdonalds)).*

3. Retrieval system, test data and queries

3.1 Retrieval system

The test system used was the *InQuery* retrieval system (Allan et al., 2000). InQuery is a probabilistic retrieval system based on the Bayesian inference net model. Queries can be presented as bag-of-words queries, or can be structured using a variety of query operators. All query keys are attached with a belief value, which is calculated by the *tf.idf* modification of the system.

InQuery's query operators are prefix operators, i.e. these precede the operand keys. The operators are marked by the hash sign #, e.g. #*sum* and #*syn*. A white space is used between the keys in queries. Therefore, *#sum(computer virus)* is a well-formed expression in the InQuery query language. The operators used in this study were #*sum* and #*syn*. For the #*sum* operator, the system computes an average weight of query keys or subqueries. The #*syn* operator treats its operand keys as synonymous instances of the same key.

3.2 Test data

The test data of this study was CLEF 2001 data (Peters, 2002) and included a test collection and a set of test topics. The collection contains some 112 000 newspaper articles from the *Los Angeles Times*. The test topic set used in this study contained 35 of the 50 CLEF 2001 topics. (See Hedlund et al., 2004, for a discussion on UTACLIR and CLEF 2001.)

3.3 Queries

A CLEF topic consists of three fields: a title, a description and a narrative. In this study, an independent translator translated the title and description fields of the topics into Afrikaans. The Afrikaans words were translated back to English by means of the UTACLIR system to yield the test queries.

The original English CLEF topics served as baseline queries, and these were run in the study to show the performance of the test queries in a monolingual sense. All words in the topics, except for words that were found in UTACLIR's target stopword list, were included in the baseline queries. A stopword list was applied for baseline queries to avoid a bias towards the test queries.

There were two types of test queries and one baseline query for each test query. The first set of queries was formulated on the basis of the title fields of the topics only, and the second set of queries on the title as well as the description fields of the topics.

The baseline queries were flat #*sum* queries, i.e. no other operator than #*sum* was applied in the baseline queries. In the test queries, the English translation equivalents of the same Afrikaans word were combined by InQuery's #*syn* operator. This method has been shown to be effective in CLIR for handling incorrect translations and other bad keys (Darwish & Oard, 2003; Pirkola, 1998; Sperer & Oard, 2000).

Below we present a sample CLEF topic title in English and Afrikaans, and a baseline and test query based on the title:

- The title of CLEF Topic 41 in English: "Pesticides in baby food"
- The Afrikaans source query: *Plaagdoders in babakos*
- The English baseline query: #sum(pesticide baby food)
- The English target query translated from the Afrikaans source query: #sum(#syn(nullstr lues die van plague plague blight infestation pest affliction vexation killer) #syn(nullstr) #syn(baby food)) – "nullstr" refers to a stopword removed from the query

The first Afrikaans key, *Plaagdoders,* is a plural compound, untranslatable as such but splittable by the source normaliser into the components *plaag* and *doder*. The component words are translated separately and the translations are enveloped into the same synonym set (the first *#syn* statement of the target query).

The second source key, *in,* is a source stopword removed from the source query and replaced by a non-matching *"nullstr"* in the second synonym set of the target query. The third Afrikaans key, *babakos*, is found in the translation dictionary, with the translation "baby food".

4. Findings

The results were evaluated as an average precision over ten recall points (10-100%) and as precision at the 10% recall point. The results are presented in Table 2. The average precision of the test queries was 13,6% (title only) and 19,4% (title and description), while baseline queries gave an average precision of 36,2% (title only) and 39,8% (title and description).

The relative performance of test queries with respect to the baseline queries was 37,6% (title only) and 48,7% (title and description). At the 10% recall point, test queries performed better with respect to baseline queries, with the relative performance figures being 41,6% (title only) and 52,0% (title and description).

Table 2: Findings of retrieval experiments

Query type (n = 35)	Average precision (%)	% of baseline	Precision at 10% recall	% of baseline
Title				
Baseline	32,6		60,1	
Test queries	13,6	37,6	25,0	41,6
Title and description				
Baseline	39,8		63,1	
Test queries	19,4	48,7	32,8	52,0

5. Analysis of the results

After the test runs, a manual word-by-word analysis (n = 607) of the results was made. The following aspects of the process and results should be noted.

5.1 Uppercase and lowercase

The effect of modifying the uppercase starting letters of keys to lowercase in the second step of the normaliser (see Section 2.4) had a positive effect on the retrieval results. There were far more words starting with an uppercase letter (because it was found at the beginning of a sentence), than there were unmatched proper nouns.

5.2 Untranslatable words

Untranslatable words were mainly found in the following categories:

- Words not found in the dictionary, mainly technical or new computer terms
- Proper nouns, with the exception of the names of countries; it was also found that transliteration is often used when forming the Afrikaans version of especially the names of countries in Eastern Europe and Asia, e.g. *Tjetsnië* for Chechnya
- Abbreviations and acronyms, e.g. *VSA* or *VN*
- Adjectives formed from the names of countries, e.g. *Rusland* (Russia) is in the dictionary, but *Russiese* (Russian) is not
- Rare inflections not solvable by LCS matching, e.g. *skip* > *skepe* (the singular and plural form of "ship")

5.3 Compounds

Compounds were split using LCS matching and the identification and removal of joining morphemes where necessary (see Section 2.4). In general, this procedure worked well. Exceptions occurred where part of the compounds were words not found in the dictionary, e.g. computer terms and proper nouns. Hyphenated compounds were mistranslated in general (20 instances).

5.4 Homonyms

Homonyms and lexical ambiguity in source and target languages are common problems in CLIR experiments. In Afrikaans, specifically, many mistranslations occurred due to stopwords that were not included in the stopword list, because they also have an alternative content-bearing sense. Examples in these experiments are *van* – "of" as a stopword, but also meaning "surname" or "family name" (27 occurrences); *deur* – "through" as a stopword, but also meaning "door" (5 occurrences); *oor* – "over" as a stopword, but also meaning "ear" (16 occurrences). In this set of test queries, none of the occurrences was used for the alternative meaning listed above. In all cases, these words occurred as stopwords and therefore should not have been translated.

Several of the individual words in the queries had very large *#syn* sets, notably *dood* ("death"), which had 126 words as synonymous meanings.

5.5 Set expressions as part of the dictionary entries

Because set expressions formed part of the original dictionary's entries, a number of Afrikaans words were included in the *#syn* sets of the English query words. For example, the entry for *aaklig* also includes examples of the use of the word in the following set expressions: *'n aaklige gesig* ("a terrible sight"), *aaklige gewoontes* ("nasty habits") and *ek voel aaklig daarvan* ("it has upset me, it made me feel sick"). All the Afrikaans words cited here would have been included in the *#syn* set if *aaklig* had to be translated. In the test data, there were 49 occurrences of query words containing Afrikaans words in the *#syn* sets.

5.6 Incidental diacritical marks

Although there were no occurrences of such marks in the test data, it is a unique characteristic of the Afrikaans language that basically any word can be emphasised by placing accent signs on the vowels in certain parts of the word. The purpose is to emphasise the context of the sentence for the reader. The accented words are, however, not found as dictionary entries. This problem was solved relatively easily for the stopwords (Section 2.3). In running text, however, these characters may appear anywhere and these words will therefore not be translatable through exact matching with the dictionary entries.

6. Discussion and recommendations

The basic research question addressed in this paper was: Based on the dictionary translation approach, how can Afrikaans information requests be translated into English queries, and how effective are such queries in retrieving documents in an English collection? The results showed that for the 35 queries based on the title and description fields of the CLEF topics, the average precision was 48,7% of that of the monolingual English queries. In CLIR literature (e.g. Hedlund et al., 2004), relative precision figures of 60% to 80% are commonly reported. We consider our results promising, given that we have developed and tested a fully automatic Afrikaans-English CLIR system and have identified several problems and suggested solutions for them. In the South African context, when compared with the results of the Zulu-English CLIR reported by Cosijn et al. (2002), the results were substantially better.

For these experiments, an electronic bilingual translation dictionary was engineered to be used for CLIR specifically. The process has not yet been optimised and there is room for improvement through further filtering of the dictionary. It was found that due to the set expressions included in the dictionary, many Afrikaans words are included in the translated *#syn* sets, which has a negative effect on the retrieval performance. It was also found that some of the dictionary entries are very long, which also impacted

negatively on the results. Word frequency analysis to establish the most common senses of these words might improve the results of future experiments.

The normaliser that was created performed quite well. Identification and normalisation of plural forms and past tense verbs were mostly correct. LCS matching to normalise words with a variety of suffixes was successful.

The process for compound splitting worked well, except in cases where individual parts of the compounds were not found in the dictionary, and in a few (rare) cases where the beginning of a proper noun was also an entry in the dictionary, e.g. *krugersdorp* was not recognised as a proper noun and was matched with *kru* ("crude"). In these experiments, the individual parts of the compounds were handled as synonyms, e.g. *tienerselfmoorde* was translated and handled as *#syn(teenage suicide)*, but subsequent experiments have shown that the results improve when handled as separate *#syn* sets, i.e. *#sum(#syn(teenage) #syn(suicide))*. Hyphenated compounds were not handled correctly. This will be corrected in a future version of the normaliser.

The stopword list was found to be adequate, except in the cases mentioned under Section 5.4. Frequency analysis should be undertaken to establish how often homographs occur that are also stopwords in a sense, and the stopword list should be modified accordingly.

7. Conclusion

The results of this first query-based CLIR experiment ever undertaken from Afrikaans to English are quite promising and an analysis shows that there is potential for significant improvement. Although standard procedures for query-based CLIR were generally followed, a number of analytical innovations were introduced which have expanded the understanding of the process. The main contributions of this research include the engineering of a bilingual translation dictionary, the compilation of an Afrikaans stopword list, and the development of a morphological analyser specifically for CLIR to normalise Afrikaans words in running text to be matched to the dictionary entries. Not only do the findings enable actual CLIR processes in South Africa, but the basic analytical process has also been expanded and the applicability of UTACLIR for other languages verified.

ACKNOWLEDGEMENTS

The InQuery search engine was provided by the Centre for Intelligent Information Retrieval at the University of Massachusetts.

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright ©1989-1992 Atro Voutilainen and Juha Heikkilä.

TWOL-R (Run-time Two-Level Program): Copyright ©Kimmo Koskenniemi and Lingsoft plc. 1983-1992.

REFERENCES

Allan, J., Connell, M.E., Croft, W.B., Feng, F-F., Fisher, D. & Li, X. 2000. *InQuery and TREC-9: The Ninth Text REtrieval Conference (TREC-9)*, Gaithersburg, MD. [Online]. http://trec.nist.gov/pubs/ trec9/t9_proceedings.html. Accessed 30 August 2004.

Cosijn, E., Bothma, T.J.D., Järvelin, K., Pirkola, A. & Nel, J.G. 2002. Metadata and cross-language information retrieval as complementary technologies to provide access to knowledge databases in indigenous languages. In: Bothma, T. & Kaniki, A. (Eds), *Progress in Library and Information Science in Southern Africa: Proceedings of the Second Biennial DISSAnet Conference (ProLISSA 2)*. Pretoria: Infuse, 397-408.

Darwish, K. & Oard, D. 2003. Probabilistic structured query methods. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 338-344.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. 2004. Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval,* 7(1/2): 99-119.

Oard, D. & Diekema, A. 1998. Cross language information retrieval. *Annual Review of Information Science and Technology*, 33: 223-256. Medford, NJ: Information Today.

Peters, C. 2002. *CLEF: Cross-Language Evaluation Forum.* [Online]. http://clef.iei.pi.cnr.it:2002/. Accessed 30 August 2004.

Pirkola, A. 1998. The effects of query structure and dictionary set-ups in dictionary-based cross-language information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 55-63.

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.

Sperer, R. & Oard, D. 2000. Structured translation for cross-language IR. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, 120-127.