Chapter 4

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

Ralph Waldo Emerson

4 Results of previous empirical studies

4.1 *Introduction*

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

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

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

The following issues will be dealt with in this chapter:

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

4.2 *A brief overview of CLEF*

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

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

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

Table 4.1 *A roadmap of CLEF*

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

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

4.3 Methods and data used in querying the database

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

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

Firstly, the English query was translated into Zulu (one pair), and this Zulu query was then matched to the database index (also in Zulu). This should be applied to all the necessary pairs. Matches could be made through various techniques, but because no suitable Zulu morphological analysers were available, a dictionary translation method had to be used (see Section 2.3.1). N-grams was used to match the Zulu search strings with the inverted index. Since there are many morphological analysers available for English, it was trivial to match the English words in natural language to the translation dictionary entries. For each English word, a number of Zulu translations were found, some of which were correct and some of which were incorrect in context of the query (owing to the ambiguity of natural language). All these words were then matched against the inverted index. Since the inverted index is not normalised, approximate string matching between the query words and indexed words was necessary. By using n-grams, this was managed quite well.

4.4 The retrieval system and test queries

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

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

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

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

N-grams are formed of the adjacent characters of words. The *classified s-gram (skipgram) matching* technique is a new n-gram matching technique that is described in more detail in Pirkola et. al., (2002a). In this technique, digrams are combined both of adjacent and non-adjacent characters of words. Digrams are classified into categories based on the number of skipped characters. Only digrams belonging to the same category are compared with one another. In Section 2.2.1.1, the problem of words not always being translated from the source language into target language was discussed. One of the most widely known methods for handling these translation errors is to pass the untranslated words to a CLIR query (the final target language query) as such. In the instance of spelling variants however, a source language form does not match the variant form in a database index, causing loss of retrieval effectiveness. Therefore, alternative methods for translation needs to be applied–this is where n-gram matching (and other approximate string matching techniques), or transliteration based on phonetic similarities between languages–comes in as a useful method to find target language spelling variants for source language words (Pirkola et al., 2001).

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

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

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

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

The digrams for the three words are as follows:

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

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

| {um, mu, un, nt, tu} ∩ {nt, tu} | / | {um, mu, un, nt, tu} ∪ {nt, tu} | =
| {nt, tu} | / | {um, mu, un, nt, tu} | = 2/5 = 0,4

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

| {um, mu, un, nt, tu} ∩ {ab, ba, an, nt, tu}| / | {um, mu, un, nt, tu} ∪ {ab, ba, an, nt, tu}| =
| {nt, tu} | / | {um, mu, un, nt, tu, ab, ba, an} | = 2/8 = 0,25

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

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

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

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

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

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

#sum(a b c ...)

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

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

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

The syn-statement is as follows:

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

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

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

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

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

As saturation normally occurs around the first three translated matches, the correct Zulu word should be identified within a set of three words about 80% of the time. Each word in the Zulu translations of the query topics was matched to the base-forms in the electronic monolingual Zulu word list. Based on the five topics analysed (as described above) it was decided that the three best matches of the skipgram technique should be used, which identifies the correct base form in 78% of all cases. For each source word, the three best matches were translated and these were to be treated as synonyms.

4.5 Zulu to English Translation

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

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

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

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

For all CLIR queries (undisambiguated, syn1, and syn2) proper names were managed differently. Borrowed proper names in Zulu are usually prefixed, with the stems being in their source language forms, e.g., *eSiberia*. The capital letters are kept in the middle of the proper names. This simplifies proper name identification and handling in CLIR. The prefixes were removed from the Zulu proper names (single and hyphenated words having a capital letter in the middle of words), and the unprefixed forms (e.g., *Siberia*), were used in the final queries. In Zulu, borrowed phrases (common nouns) are usually indicated by applying inverted commas, e.g., "computer virus". As such, both unprefixed proper names and borrowed phrases were passed to the final queries. The number of queries containing these two expression types was 34 and 6 respectively (some queries contained both). The number of queries that did not contain either one of these was 13 (see Table 4.3). Because proper names in Zulu are identical to their English forms, removing the Zulu prefix from the proper name resulted in satisfactory retrieval performance of the syn-queries. However, it is not always the case that proper names are identical in Zulu and English respectively. Proper name queries also contained many mistranslated keys, but owing to the syn-structure it performed quite well.

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

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

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

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

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

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

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

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

Approximate string matching provided relatively good results, but several types of problems were experienced (e.g. inflected word forms and paraphrased translations). The correct base-forms were not always top-ranked, and this caused ambiguity in the translation process. It was found, that although paraphrasing may give some good keys, the syn-structure does not help in these instances. The latter happened because of the disambiguation effect of the syn-structure ((important keys occur as single keys (outside the syn-statements)). Also see Section 4.4 for more detail on the syn-structure of the queries. This implies that more weight is assigned to single keywords than to mismatched keys. The mechanical matching of running text to dictionary entries through approximate string matching works relatively well, but at a conceptual level there are serious problems. This includes strings that are incorrectly recognized (Lopresti and Wilfong, 1999). For example, the bilabial implosive–previously written as *б*–is represented by b, as in *ubaba, ubudoda*. It is a very common error for systems to identify the *б* for a 6 in the written form. Unfortunately, the pronunciation of the words cannot be taken into consideration, as it sounds the same in almost all instances. One then needs to examine the grammatical analysis of the word to make a correct identification.

4.6 Chapter Synopsis

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

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

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

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

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

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

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