Authors:	Pirkola Ari, Puolamäki Deniz, Järvelin Kalervo
Name of article:	Applying query structuring in cross-language retrieval
Year of publication:	2003
Name of journal:	Information Processing & Management
Volume:	39
Number of issue:	3
Pages:	391-402
ISSN:	0306-4573
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URN: http://urn.fi/urn:nbn:uta-3-745
DOI: http://dx.doi.org/10.1016/S0306-4573(02)00091-2

Applying Query Structuring in Cross-Language Retrieval

Ari Pirkola, Deniz Puolamäki and Kalervo Järvelin
Dept. of Information Studies
University of Tampere

Preprint of:
Pirkola, A. & Puolamäki, D. & Järvelin, K. (2002). Applying query structuring in cross-language retrieval. Information Processing & Management 38(xxx): xxx-xxx, in press.

Applying Query Structuring in Cross-Language Retrieval

Abstract
We will explore various ways to apply query structuring in cross-language information retrieval (CLIR). In the fist test, English queries were translated into Finnish using an electronic dictionary, and were run in a Finnish newspaper database of 55, 000 articles. Queries were structured by combining the Finnish translation equivalents of the same English query key using the syn-operator of the InQuery retrieval system. Structured queries performed markedly better than unstructured queries. Second, the effects of compound-based structuring using a proximity operator for the translation equivalents of source language compound components were tested. The method was not useful in syn-based queries but resulted in decrease in retrieval success. Proper names are often non-identical spelling variants in different languages. This allows n-gram based translation of names not included in a dictionary. In the third test, a query structuring method where the Boolean and-operator was used to assign more weight to keys translated through n-gram matching gave good results.

Keywords: Compound word processing, Cross-language information retrieval, N-gam matching, Proper name searching, Structured queries

1. Introduction

Cross-language information retrieval (CLIR) refers to an information retrieval task where the language of queries is other than that of the retrieved documents. A user may present a query in his or her native language, and in response the system retrieves documents in another language. For an overview of CLIR approaches, see (Hull and Greffenstette, 1996; Oard and Diekema, 1998; Pirkola et al., 2001). In dictionary-based cross-language retrieval, queries are translated by means of electronic dictionaries by replacing source language query keys with their target language equivalents.

In this paper we will study three different conditions in which *query structuring* could be applied in CLIR to improve retrieval effectiveness. The basic idea of query structuring is to group query keys and to use query operators in such a way that more weight is being assigned to important or correct keys than the other keys. To implement this, there has to be some indicator available to show the probability of a key being a good or bad key. Two factors implicitly present in dictionaries are assumed to indicate key goodness. First, it is assumed that important keys often have 1-2 translation equivalents while less important have several equivalents. Second, proper names often are important keys but general dictionaries do not generally include proper names. Therefore the absence or presence of a key in a dictionary is regarded as another indicator.

Pirkola (1998) and Pirkola et al. (2002) studied the effects of query structure on CLIR performance in cross-language retrieval where the source language was Finnish and the target language English (*Fin-Eng* CLIR). Queries were structured by grouping together the English equivalents of the same Finnish query key by means of the *syn*-operator of the InQuery retrieval system. The syn-operator

treats its operand query keys as synonyms, i.e., instances of the same query key. Structuring caused significant improvements in retrieval performance for long queries (formed on the basis of the title and description fields of TREC topics) and modest improvements for short queries (title fields) with respect to the performance of unstructured (undisambiguated) queries. The relative performance improvements were 107% for long queries and 18% for short queries (Pirkola et al., 2002). These figures show that in *Fin-Eng* CLIR syn-based structuring is very effective for queries characterized by the abundance of mistranslated and other bad keys and helpful for queries which contain few bad keys.

The query structuring technique shown to be effective in *Fin-Eng* text retrieval could be implemented readily in operational retrieval systems. However, it is possible that the specific linguistic features of Finnish as a source language or English as a target language contributed to the good performance of the structured queries. Thus, the effectiveness of the method may depend on the languages of a CLIR system as well as the direction of translation. In this paper we will explore whether the method is useful also in *Eng -Fin* text retrieval, i.e., the case where translations are done in an opposite direction with respect to those done by Pirkola (1998) and Pirkola et al. (2002). Thus, we will seek evidence that the syn-structuring technique holds generally, irrespective of a language pair used in a CLIR study or a system.

It is known that languages differ considerably in linguistic features, such as the frequency of lexical ambiguity (Ullman, 1967) and morphological features (Greenberg, 1960; Pirkola, 2001). English and Finnish are different types of languages, particularly in morphology. In English grammatical relations are indicated mainly by prepositions while Finnish typically uses a grammatical case. In Finnish, there are altogether 14 features in the category of case (Karlsson, 1987). Therefore, the number of word forms that a Finnish noun may take is very high, theoretically 2200 forms (Karlsson, 1983). Inflection has depressing effect on CLIR effectiveness, but it is hard to estimate in which case, *Fin-Eng* or *Eng -Fin*, the problem is more severe. In *Fin-Eng* retrieval, inflection causes difficulties especially in query processing whereas in *Eng -Fin* retrieval troubles occur in indexing.

In Finnish multiword expressions are typically compound words, in English they are often phrases[1]. Thus, a Finnish compound word is often translated as a noun phrase in English. From the IR and CLIR perspectives, a compound word is a more convenient type of expression than a phrase, because compound decomposition is easier than phrase identification. In this respect *Fin-Eng* retrieval may be easier than *Eng -Fin* retrieval. Finnish compounds can be split effectively into component words by a dictionary-based morphological analyzer. The English equivalents of the components can be combined by a proximity operator in CLIR queries(the final target language queries).

Second, we will test the effects of *compound-based structuring* in CLIR. It refers to a query structuring method where a proximity operator (Inquery's uwn-operator) is used to combine the translation equivalents that correspond to the first part of a source language compound to the equivalents that correspond to the second part of the compound. This method has been used in some studies in which the effects of the syn-operator have been tested (Gollins, 2000; Hedlund et al., 2001; Pirkola, 1998). However, the effects of the uwn-operator have not been tested; it is not clear what is the contribution of compound-based structuring to the effectiveness of syn-based queries.

[1] *Compound word* is defined here as a multiword expression in which the components are written together. *Phrase* is defined as a multiword expression in which the components are written separately.

Therefore we will test in this study whether syn-structured queries benefit from using the proximity operator for the translation equivalents of compound components.

Third, we will examine query structuring in combination with *n-gram based translation* (matching) of words not included in a dictionary. General translation dictionaries may include some proper names, such as the names of capital cities and countries. The vast majority, however, is not covered. Particularly, translation dictionaries do not contain personal names. Typically untranslatable keys are used in their original form in a CLIR query. Proper names in different languages often are non-identical spelling variants, e.g., *Brussels* (in English) and *Bryssel* (in many languages). In these cases a source language name does not match its variant form in the target database index, causing loss of retrieval effectiveness.

However, the fact that proper names often are spelling variants of each other allows the use of n-gram matching or some other fuzzy matching (approximate string matching) technique for proper name translation. N-gram matching is a language independent means to recognize words whose character strings resemble each other (Angell et al., 1983; Hall and Dowling, 1980; Robertson and Willett, 1998). Query keys and the words of documents are decomposed into n-grams, i.e., into substrings of length *n.* The degree of similarity between query keys and index terms can then be computed by comparing their n-gram sets. N-gram matching has been reported to be an effective method to find monolingual name spelling variants and spelling error forms (Pfeifer et al., 1996; Robertson and Willett, 1998; Zobel et al., 1995) as well as cross-lingual name spelling variants (Pirkola et al., 2002). In this paper, we will investigate whether query structuring in combination with n-gram translation will improve retrieval effectiveness.

2. Methods and data

In all our tests, the test system was the *InQuery* retrieval system (Allan et al., 2000). InQuery is a probabilistic information retrieval system based on Bayesian inference net model. Queries can be presented as bag of word queries, or they can be structured through a variety of query operators.

In all tests of this study, the source language words were translated to a target language by an electronic MOT *Eng-Fin-Eng* dictionary by Kielikone Plc. The dictionary contains 165,000 entry words, of which 65,000 are Finnish and 100,000 English words. In all the tests, the words that were not found as entry words in the dictionary were sent unchanged to the CLIR queries. Such expressions were proper names, acronyms, and English words not found in the dictionary.

2.1. *Syn-based query structuring (English–Finnish CLIR)*

The Finnish test collection contained around 55.000 articles published in three Finnish newspapers in 1988-1992. Our test environment provides 35 test requests (in Finnish) for which the relevance of 16 000 articles is known (Kekäläinen and Järvelin, 1999). 20 of the 35 requests were used as test queries in this study. The query set of 20 queries is small, and the findings of this test should be considered together with the recent findings of syn-based query structuring (see the Discussion section).

The database index contained the inflected word forms of the Finnish documents in their normalized base forms. Compound words were stored as normalized compounds as well as their normalized component words. Normalization and compound splitting were done by a dictionary-based morphological analyzer Fintwol. The queries were natural sentences. For this study, the inflected word forms of the Finnish (*baseline*) queries were similarly normalized into their base

forms. The compounds of Finnish queries were split and the normalized component words of compounds were combined by an uw3-operator in queries. This was done to get a strong baseline. The use of the uw3-operator is useful in monolingual (Finnish) retrieval. The uwn (unordered window) operator allows for free word order in the proximity statements. The window size *n* refers to the number of spaces between words in the text.

The Finnish queries were translated into English by the author of this paper. The translations were checked by a colleague whose native language is English. Human translation was done to get test queries that are comparable to the original Finnish queries. Strict rules were followed in translation.

The test query types were as follows. The query operators used in the queries are presented in parentheses. Sample queries are presented later in this section.

1. Finnish queries, i.e., baseline for the CLIR queries of steps 2-3 (#sum, #uw3)

2. Unstructured CLIR queries (#sum)

3. Structured CLIR queries (#sum, #syn for translation equivalents derived from the same English word)

The query keys of (1) *baseline queries* were the same words in normalized forms that occurred in the Finnish requests. The keys were combined by the *sum*-operator of the InQuery retrieval system. For the sum-operator, the system computes an average weight of query key weights. In (2) *unstructured queries*, the translated keys were combined by the sum-operator. In (3) *structured queries*, the Finnish equivalents of the same English key were grouped together by the *syn*-operator. This was done in all those cases where an English key had more than one equivalent. The syn-statements and single words (in the case of just one translation equivalent) were combined by the sum-operator to give the final CLIR query.

The queries below illustrate the differences between baseline queries and CLIR queries, as well as the differences between structured and unstructured CLIR queries.

1. Baseline query (the original Finnish query)
#sum(helsinki tapaaminen george bush mihail gorbatshov syyskuu asia käsitellä päätös sopimus tehdä neuvotella)

English source words (request words in base form, excluding stop words)
helsinki, summit, george, bush, mikhail, gorbachev, september, issue, discuss, decision, agreement, made, negotiation.

2. Unstructured CLIR query
#sum(helsinki harja huippu huippukokous huipputapaaminen lakipiste vuorenhuippu george bush mikhail gorbachev syyskuu emissio emittoida jakaa julkaista jälkeläinen kysymys laskea painos keskustella neuvotella pohdiskella pohtia päätäntävalta päätöksenteko päätös ratkaisu tuomio kauppa kongruenssi liitto myöntymys myötävaikutus sopimus yhtäpitävyys made tekoinen valmistaa neuvottelu välitys)

3. Structured CLIR query
#sum(helsinki #syn(harja huippu huippukokous huipputapaaminen lakipiste vuorenhuippu) george bush mikhail gorbachev syyskuu #syn(emissio emittoida jakaa julkaista jälkeläinen kysymys laskea painos) #syn(keskustella neuvotella pohdiskella pohtia) #syn(päätäntävalta päätöksenteko päätös ratkaisu tuomio) #syn(kauppa kongruenssi liitto myöntymys myötävaikutus sopimus yhtäpitävyys) #syn(made tekoinen valmistaa) #syn(neuvottelu välitys))

In the latter example the *syn*-operator indicates which Finnish words were translation equivalents of

the same English word.

Generally, the syn-based queries are formed as follows. For a source language query containing the keys A, B, and C, the target language syn-query is:

#sum(#syn(a_1... a_n) #syn(b_1... b_m) #syn(c_1... c_k))

where a_1... a_n, b_1... b_m, c_1... c_k stand for the translation equivalents of the keys A, B, and C, respectively.

2.2. *Compound-based structuring (Finnish-English CLIR)*

The test document collection was a TREC collection and contained 515,000 documents. The test queries used in the tests of compound-based structuring were formed from the TREC topics 76-150. The words of the title and description fields of the topics were used as query keys. A professional translator translated the topics (title and description fields) into Finnish according to the guidelines provided by CLEF (Peters, 2000). The topics that contained at least one compound word (n=47) were selected as test queries for this test. The Finnish words were retranslated to English using an automatic query translation and construction system developed in the Information Retrieval Laboratory at the University of Tampere (the system is called UTACLIR). The MOT *Eng-Fin-Eng* dictionary is part of the system. The language and query processing components of the UTACLIR system are described in Hedlund et al. (2001) and Pirkola et al. (2001).

Three query types were studied in the test. For a source language query including a compound AB and a single word C the translation equivalents are marked as a_1... a_n, b_1... b_m, and c_1... c_k, respectively.

(1) #sum(#syn(a_1...a_n) #syn(b_1...b_m) #syn(c_1... c_k))

(2) #sum(#syn(#uw3(a_1 b_1) … #uw3(a_n b_m)) #syn(c_1... c_k))

(3) #sum(#syn(#uw3(a_1 b_1) … #uw3(a_n b_m) a_1...a_n b_1...b_m) #syn(c_1... c_k))

In the latter case the component words are repeated in the syn-statement. This may be useful for compounds where just one of the component words is semantically transparent.

2.3. *N-gram matching (Finnish-English CLIR)*

The n-gram tests were performed using CLEF's (Peters, 2000) LA Times document collection as a test collection, which was indexed using the Engtwol morphological analyzer. The collection contains some 112,000 documents. Of the CLEF'2001 topics (n=50) we selected for this test as test queries those topics (n=26) which contained untranslatable keys (the words not found in a translation dictionary or the dictionary of a morphological analyzer). The title and description fields of the topics were translated by a professional translator into Finnish according to the guidelines provided by CLEF. The Finnish words were translated back to English by means of UTACLIR.

Proper names and other words not contained in the translation dictionary were translated by an n-gram matching technique included in UTACLIR. Query keys and index terms were split into n-grams combined both of adjacent charcters of words as well as *non-adjacent characters separated*

by one character in the words. The method is described in Pirkola et al. (2002).

We tested the effects of syn-structure for queries where two, four and six highest ranked (HR) keys in the result list of n-gram matching (here we call them *HR n-gram keys*) were selected as keys for the CLIR queries. In the database index, those words that the morphological analyzer Engtwol did not recognize were indexed in a separate index and were marked by the @-mark. In the case of two HR n-gram keys, the other key was the HR key from the index of recognized words and the other one the HR key from the index of the unrecognized words. Correspondingly, in the case of 4 (6) HR n-gram keys two (three) were recognized words and two (three) unrecognized words.

Our assumption in the beginning of the study was that syn-structure in combination with n-gram translation is useful. However, it turned out that the best results were achieved using two HR n-gram keys without using syn-stucture. This finding suggests that the words translated through n-gram matching were the most important keys of queries; as an important key is embedded in a syn-statement its relative weight is decreased, and this results in the drop in query performance (see the Discussion section). Therefore, in line with our earlier findings in Pirkola and Järvelin (2001) (see below) we tested the use of the Boolean and-operator (Inquery's *band*-operator) to assign the n-gram keys more weight than the other keys of a query. All the argument keys of the band-operator must occur in a document in order for the operator to contribute to the weight computed for that document. The following structure was used. As shown, the n-gram keys are combined with the other keys by the band-operator, and in addition the syn-structure is used:

#sum(#band(key_1 key_2) #band(key_1 key_3) … #band(key_1 key_n) … #band(key_2 key_3) … #band(key_2 key_n) [syn-strucre])

Key_1 and key_2 denote the HR n-gram keys and key_3… key_n the keys translated by a dictionary.

In the band-structured queries the syn-structure is used to relax the strict conjunction restrictions of the Boolean operator. Sometimes the HR n-gram key is a bad key. In these cases the syn-structured subquery is important for succesful retrieval.

For example, the Finnish query 50 contained an untranslatable key *Chiapasissa* (an inflectional word form of the name *Chiapas*). The two HR n-gram keys and the final band/syn-query is as follows:

#sum(#band(chiapas @chiapis) #band(chiapas insurrection) #band(chiapas mutiny) #band(chiapas rebellion) #band(chiapas revolt) #band(chiapas rising) #band(chiapas uprising) #band(chiapas american) #band(chiapas indian) #band(chiapas national) #band(chiapas mexico) #band(@chiapis insurrection) #band(@chiapis mutiny) #band(@chiapis rebellion) #band(@chiapis revolt) #band(@chiapis rising) #band(@chiapis uprising) #band(@chiapis american) #band(@chiapis indian) #band(@chiapis national) #band(@chiapis mexico) #syn(insurrection mutiny rebellion revolt rising uprising) #syn(american indian) #syn(national revolt rising) mexico)

In Pirkola and Järvelin (2001) we showed that the application of the band-operator in this way in monolingual retrieval will improve retrieval performance. We showed that in most (TREC) topics 1-2 words are crucial in terms of retrieval performance. If the most important key is removed from a query, the performance of the query will drop dramatically. The most important keys often are proper names. For example, in the topic *(Document will report on) the negotiating process leading*

to an end to the Nicaraguan civil war, the removal of the name *Nicaraguan* will ruin query performance.

3. Findings

The results were evaluated as (1) average precision over ten recall points (10-100%), as (2) precision-recall graphs, and (3) as precison at 10% recall.

The results of syn-based structuring are presented in Table 1 and Figure 1. As shown in Table 1, the average precision of structured queries is 27.4% while unstructured queries give the precision figure of 18.8%. The relative improvement percentage due to structuring is 45.7% (column 3). As shown in column 4, the relative performance of CLIR queries with respect to the baseline queries is 77.0% (structured queries) and 52.8% (unstructured queries).

Figure 1 shows precision-recall curves for CLIR and baseline queries. As can be seen, structured queries perform markedly better than unstructured queries at all recall levels, but fall below baseline queries.

Table 1. The effects of syn-based structuring (Eng-Fin CLIR)

Query Type *Number of queries, N=20*	*Avg. Precision*	*% Change Str. vs. Unstr.*	*Perf. in relation to baseline perf.*
Structured queries	27,4	45,7	77,0
Unstructured queries	18,8	-	52,8
Baseline (Finnish) queries	35,6	-	100,0

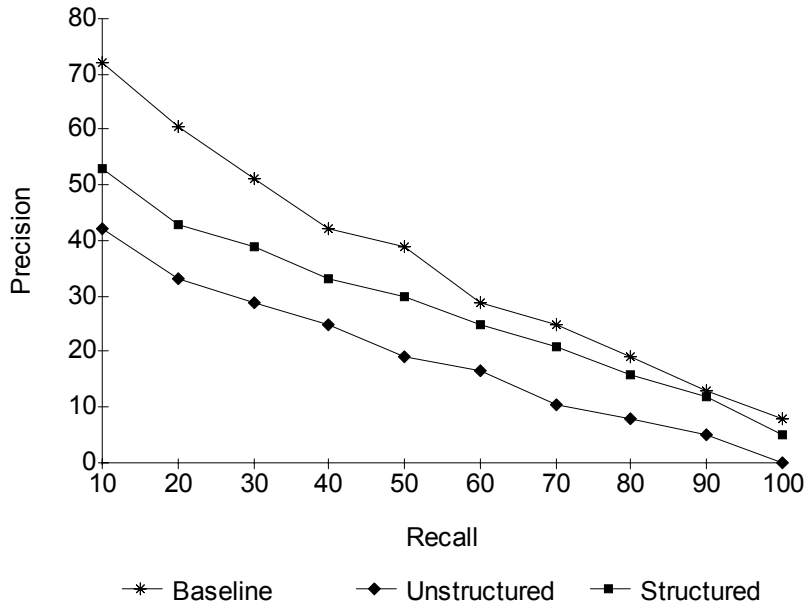

Figure 1. Precision-recall curves for CLIR and baseline queries

The findings of compound-based structuring are shown in Table 2. The absolute performance figures for short queries not tested in this study are better than for long queries (Pirkola et al., 2002). As can be seen, syn-queries without the uw3-operator perform better that the syn-queries where the equivalents of compound word components are joined by the uw3-operator. The query type where the keys are repeated in the syn-statements is somewhat better than the other uw3-query type.

Table 2. The effects of compound-based structuring (Fin-Eng)

Query Type *Number of queries, N=47*	*Avg. Precision*	*Precision at 10% Recall*
Syn-based structuring	8,0	20,3
Syn + uw3	6,3	16,8
Syn + uw3 + repeated keys	5,9	15,9

Table 3 presents the results of n-gram matching. In all cases n-gram based queries perform better than the queries in which n-gram matching is not applied. Syn-structure is useful in the case of 6 HR n-gram keys included in a query. However, in the cases of 2 and 4 HR keys (avg. precision), the unstructured queries perform better than the syn-structured queries. Of all the query types tested the band/syn-structured queries give the best results.

Table 3. The effects of query structure in n-gram translatation (Fin-Eng)

Query Type *Number of queries, N=26*	*Avg. Precision*	*Precision at 10% Recall*
No n-gram matching	34,9	52,5
6 HR n-gram keys		
Syn-structure in the n-gram set	41,7	59,2
No syn-structure in the n-gram set	39,4	58,1
4 HR n-gram keys		
Syn-structure in the n-gram set	40,8	57,7
No syn-structure in the n-gram set	40,3	58,9
2 HR n-gram keys		
Syn-structure in the n-gram set	41,3	59,4
No syn-structure in the n-gram set	44,1	63,7
Band/syn-structure	46,8	70,5

4. EMBEDDiscussion

In summary, substantial CLIR performance improvement was obtained by applying a structuring method in which the Finnish equivalents of the same English key were treated as synonyms and were combined by the syn-operator of the InQuery retrieval system. This finding is consistent with the findings by Pirkola (1998) and Pirkola et al. (2002) which indicated that in *Fin-Eng* retrieval the performance of structured queries was substantially better than that of unstructured queries, as well as the findings of Ballesteros and Croft (1998) who tested syn-based structuring in *English-Spanish* retrieval. In fact, after this study was completed new findings were published which show that syn-structured queries are effective for various language pairs (Gollins, 2000; Hedlund, et al., 2001; Meng, et al., 2000; Oard and Wang, 2001; Sperer and Oard, 2000). For example, Sperer and Oard (2000) studied Chinese-English text retrieval, and tested structured queries formed on the basis of a structured dictionary. Various ways to combine target language keys, including the use of InQuery's syn-operator were tested. The positive effects of syn-based structuring were significant. Thus, the findings of the present study and other studies show that the method is effective for different types of languages.

There are two important ways by which syn-based structuring improves CLIR performance. First, important keys often have 1-2 translations only, and have relatively more weight in structured than in unstructured CLIR queries. Second, syn-based structuring alters (relatively) tf.idf weights of keys in a query. In unstructured queries mistranslated and other bad keys with low document frequency may deteriorate query performance. In syn-structured queries the bad keys are downweighted because of the aggregate document frequency (Sperer and Oard, 2000).

The findings showed that syn-queries performed better than the combined syn/uwn queries. An important point in this is that one has to make difference between *the disambiguation effect of compound-based structuring* and *phrase-based searching*. The disambiguation effect refers to the fact that normally a combination of two mistranslated keys does not make any sense. Therefore, the proximity combination method applied in compound-based structuring probably has a clear disambiguation effect. On the other hand, many studies on monolingual retrieval which have used sophisticated linguistic analysis or statistical methods have shown that phrase-based searching does not improve retrieval performance, or that improvements are just small (Buckley, et al., 1995; Mitra, et al., 1997). It should be noted that this monolingual component is involved in CLIR. One should also note that there are different types of compounds regarding semantic transparency. It is likely that different types of compounds should be handled in different ways for improved CLIR performance. This issue needs further investigation.

The experiments on query structuring in combination with n-gram translation show that the best retrieval performance is achieved using two HR n-gram keys. In this case, syn-structure is not useful but band/syn-structure results in clear performance improvements. Our conclusion is that is due to the fact that proper names (many of which are untranslatable words) often are the most important keys of queries.

The query translation approach applied in this research can be criticized for its query time efficiency. Computing the synonym set weight for each source language key requires a union of the postings lists of the target translation equivalents be taken. As pointed out by Oard and Ertunc (2002), efficiency can be improved by indexing-time translation, from document language terms to their query language equivalents. This may be a viable strategy if there are only one or very few query source languages to a document collection. However, as Oard and Ertunc (2002) note, query

time translation easily supports a broad range of source languages.

5. Conclusions

We found in this paper that, as in *Fin-Eng* text retrieval, syn-based query structuring is effective in *Eng -Fin* text retrieval, causing substantial improvements in retrieval performance. This finding provides further evidence that the syn-queries are effective irrespective of languages of a CLIR system. Characteristic linguistic features of specific languages may, however, affect the relative effectiveness of the technique.

Syn/uwn queries are not as effective as syn-queries. Thus, compound-based structuring is not helpful in syn-based queries. However, there are different types of compound words regarding semantic transparency. If a compound is sematically opaque it seems clear translating the component words separately and applying a proximity operator for the equivalents does not yield good results. The question of semantic transparency needs further investigation.

Many proper names are untranslatable due to limited coverage of translation dictionaries. They may also stay unnormalized in database index construction or query processing because of not being included in the lexicon for morphological analysis. For these cases, n-gram translation is helpful. N-gram matching is a language independent means to recognize translation equivalents and is suited for many language pairs used in CLIR. N-gram keys are often proper names and the most important keys of queries. We showed in this study that assigning them more weight than the other keys of a query will result in the improvement in retrieval success. We regard this as an interesting finding that needs to be studied more.

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), Gaithesburg, MD.
Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Angell, R., Freund, G. & Willet, P. (1983). Automatic spelling correction system using a trigram similarity measure. *Information Processing & Management*, 19 (4), 255-261.

Ballesteros, L. & Croft, W.B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 64-71). Melbourne, Australia.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART: TREC-4. *The Fourth Text REtrieval Conference (TREC-4)*. Gaithesburg, MD.
Available at: http://trec.nist.gov/pubs/trec4/t4_proceedings.html

Gollins, T.J. (2000). Dictionary based transitive cross-language information retrieval using lexical triangulation. University of Sheffield. Master of Science Thesis.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26, 178-194.

Hall, P. & Dowling, G. (1980). Approximate string matching. *Computing Surveys*, 12 (4), 381-402.

Hedlund, T., Keskustalo, H., Pirkola, A., Sepponen, M. & Järvelin, K., (2001). Bilingual tests with Swedish, Finnish and German queries: dealing with morphology, compound words and query structure. Carol Peters (ed.), *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science* 2069, Springer 2001, pp. 211-225.

Hull, D. (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47 (1), 70-84.

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,* pp. 49-57. Zürich, Switzerland.

Karlsson, F. (1983). *Suomen kielen äänne- ja muotorakenne*. [Phonological and morphological structures in Finnish]. Porvoo: WSOY. [In Finnish]

Karlsson, F. (1987). *A Finnish grammar*. Porvoo: WSOY.

Kekäläinen, J. & Järvelin, K. (1999). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information retrieval*, 1 (4), 329-344.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 191-202. Pittsburg, PA,

Meng, H., Chen, B., Grams, E., Khudanpur, S., Lo, W-K., Levow, G-A, Oard, D., Schone, B., Tang, K., Wang, H-M., & Wang, J.Q. (2000). Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval. *HLT 2001, Human Language Technology Confererence*, March 18-21, 2001, San Diego, California.

Mitra, M., Buckley, C., Singhal, A. & Cardie, C. (1997). An analysis of statistical and syntactic phrases. *Proceedings of RIAO'97, Computer Assisted Information Searching on the Internet*, Montreal, Canada, pp., 200-214.

Oard, D. & Diekema, A. (1998). Cross language information retrieval. *Annual Review of Information Science and Technology,* 33, pp. 223-256.

Oard, D. & Wang, J. (2001). NTCIR-2 experiments at Maryland: Comparing structured queries and balanced translation. *The Second NTCIR Workshop*, March 7-9, Tokyo, Japan

Oard, D. & Ertunc, F. (2002). Translation-based indexing for cross-language retrieval. In: Crestani, F. & Girolami, M. & van Rijsbergen, C.J. (Eds.). *Advances in Information Retrieval, Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, Glasgow*, UK, March 25-27, 2002.

Peters, C. (2000). CLEF - Cross-Language Evaluation Forum. http://www.iei.pi.cnr.it/DELOS/CLEF

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

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

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation* 57 (3), 330-348.

Pirkola, A. & Järvelin, K. (2001). Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology*, 52(7), 575 -583.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A-P. & Järvelin, K. (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2). http://InformationR.net/ir/7-2/infres72.html.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

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

Sperer, R. & Oard, D.W. (2000). Structured translation for cross-language IR. In Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120-127. Athens, Greece.

Ullman, S. (1967). *Semantics: an introduction to the science of meaning*. Oxford.

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

Acknowledgments

The *Inquery* search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.