

# Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002

Turid Hedlund, Eija Airio, Heikki Keskustalo, Raija Lehtokangas, Ari Pirkola, Kalervo Järvelin

[turid.hedlund@shh.fi](mailto:turid.hedlund@shh.fi), [eija.airio@uta.fi](mailto:eija.airio@uta.fi), [heikki.keskustalo@uta.fi](mailto:heikki.keskustalo@uta.fi), [raija.lehtokangas@uta.fi](mailto:raija.lehtokangas@uta.fi),  
[pirkola@tukki.jyu.fi](mailto:pirkola@tukki.jyu.fi), [kalervo.jarvelin@uta.fi](mailto:kalervo.jarvelin@uta.fi)

Department of Information Studies  
University of Tampere, Finland

## **Abstract**

In this study the basic framework and performance analysis results are presented for the three year long development process of the dictionary-based UTACLIR system. The tests expand from bilingual CLIR for three language pairs Swedish, Finnish and German to English, to six language pairs, from English to French, German, Spanish, Italian, Dutch and Finnish, and from bilingual to multilingual. In addition, transitive translation tests are reported. The development process of the UTACLIR query translation system will be regarded from the point of view of a learning process. The contribution of the individual components, the effectiveness of compound handling, proper name matching and structuring of queries are analyzed. The results and the fault analysis have been valuable in the development process. Overall the results indicate that the process is robust and can be extended to other languages. The individual effects of the different components are in general positive. However, performance also depends on the topic set and the number of compounds and proper names in the topic, and to some extent on the source and target language. The dictionaries used affect the performance significantly.

## **1 Introduction**

Cross-language information retrieval (CLIR) deals with the problem of presenting an information retrieval task in one language and retrieving documents in one or several other languages. The process is bilingual when dealing with a language pair, i.e., one source language (e.g., Finnish) and one target or document language (e.g., English). In multilingual information retrieval the target collection is multilingual, and topics are expressed in one language. The process is transitive if an intermediate language is used to provide a means of translation, and, specifically, triangulated when two different translation routes are combined (Gollins & Sanderson 2001a).

The main methods for cross-language information retrieval are presented in overviews by Oard & Diekema (1998), and Pirkola et al. (2001). The methods based on the translation of queries are classified as 1) dictionary-based translation (Hull & Grefenstette 1996) 2) machine translation (Davis & Ogden 1997;

Gachot, Lang & Yang 1998) and 3) methods using parallel corpora (Sheridan, Ballerini & Schäuble 1998). The present study uses dictionary-based translation of source language queries, the documents themselves are not translated.

The *main problems* reported in direct dictionary-based cross-language information retrieval are 1) the problems of inflection 2) translation ambiguity 3) compounds and phrases and their handling 4) proper names and other untranslatable words and their translation and 5) lack of structuring (Ballesteros & Croft 1997; 1998; Hedlund 2002; Hedlund et al. 2001a; 2001b; Pirkola 1998; 1999; Pirkola et al. 2001; Sperer & Oard 2000).

*Multilingual* dictionary-based information retrieval has the same problems as presented above and in addition a merging problem. There are two approaches to solve the merging problem: 1) to build separate indexes for all the target languages, to perform retrieval from these indexes separately and finally merge the result lists, or 2) to build one merged index for all the collection languages (Nie & Jin 2002).

All the problems mentioned above, translation ambiguity in particular, occur with *transitive cross-language information retrieval*, because of the additional translation phases needed: every time a translation is performed, ambiguity is likely to increase, introducing irrelevant words into the query.

For European languages the CLEF evaluation forum provides a possibility to test cross-language system performance in a TREC-like environment<sup>1</sup>. In this study we are reporting learning experiences from the development of an automated query translation system UTACLIR developed as part of the CLEF evaluation campaign 2000-2002. The system, described in Section 2, was already initially a system based on unified principles but applied to the specifics of each language. For CLEF 2002 the process was redesigned (Hedlund et al. 2002a; Keskustalo et al. 2002; Airio et al. 2002). In its present form, it is an extendable query translation system capable of performing query translation between several source and target language pairs, using external resources like morphological normalizers, stemmers, dictionaries and stop word lists.

When developing the process, the focus has been on the following issues with respect to both source and target language words: 1) word form normalization 2) compound word handling 3) proper name handling 4) stop word removal and 5) structuring of queries. In Section 2 the process is presented. The evaluation is essential for the further development of the system. In Section 3 the aim is to get a broad and more complete view of the robustness of the whole individual process in development, by using a large test topic set in three source languages. The contribution of the components of the processes is discussed in Sections, 4, 5 and 6. Source language components, the compound handling process and the n-gram based matching

---

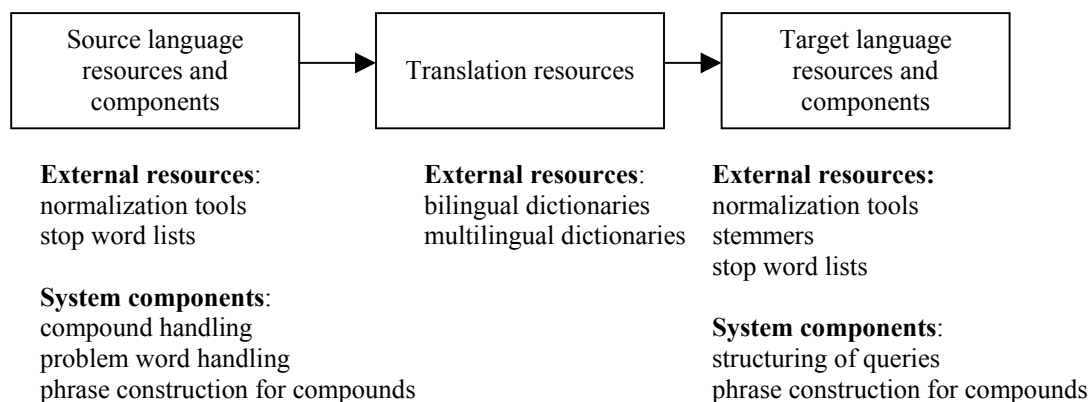
<sup>1</sup> TREC, Text Retrieval Conference, see: <http://trec.nist.gov/>

process for proper names are evaluated in Section 4. In Section 5 we discuss the impact of dictionaries and in Section 6 target language components, structuring of queries and phrase-based structuring. The multilingual process is tested in Section 7, and the transitive translation in Section 8. Discussion and concluding remarks will be found in Section 9.

## 2 The UTACLIR approach

### 2.1 System overview

The query construction framework is an automated dictionary-based process where we consider specific linguistic features of source and target language words, these include morphology and word forms not recognized by the available morphological normalizers and untranslatable words. The process has been developed for three evaluation campaigns CLEF 2000, CLEF 2001 and CLEF 2002, and therefore changes have been made throughout the query translation process. In the following we refer to the processes as Process-2000, Process-2001 and Process-2002. The processes are described in detail in the CLEF proceedings (Hedlund et al. 2001b; 2002b; Airio et al. 2002). For the CLEF 2002 evaluation campaign a beta-version of redesigned UTACLIR was in use. Many features were still lacking<sup>2</sup>. Figure 1 gives an overview of the external resources and main system components of UTACLIR.



**Figure 1 UTACLIR external resources and main system components**

The UTACLIR framework supports the use of external language resources. For each source language we utilize morphological analyzers for word form normalization. A rich inflectional morphology requires word

form normalization in order to match source keys with dictionary entries. Stop word lists are used after the normalization of words to base forms. This is important when dealing with highly inflectional source languages since the stop word lists need not contain all the inflected forms of stop words. In the present implementation, all source keys recognized as stop words are removed first. The stop word list in English was created on the basis of InQuery's default stop list for English. Finnish, Swedish and German stop word lists were created on the basis of the English stop word list. The French stop word list was granted by Université de Provence, the Italian by University of Alberta, the Dutch by University of Twente. The Spanish stop word list was InQuery's default stop list for Spanish.

The indexing of the document files in the target language is based on word form normalization using morphological analysis programs. The same analyzers are also used for normalization of topic words (Morphological analyzers: SWETWOL, FINTWOL, GERTWOL and ENGTWOL by Lingsoft plc. Finland, <http://www.lingsoft.fi>). For the indexing of the Spanish, French, Italian and Dutch document files, stemmers are used. For normalizing Spanish and French we utilize stemmers by Zprise, for Italian a stemmer by University of Neuchatel, and for Dutch a stemmer by University of Utrecht.

Machine-readable dictionaries are used for translation of source language words to the target language or in the case of transitive translations to a pivot language. In this paper we discuss dictionaries as part of evaluation.

As a test system we used the InQuery retrieval system by Center for Intelligent Information Retrieval at the University of Massachusetts. (<http://ciir.cs.umass.edu/>)

In the following sections we will assess the development stages of the process by looking at the evaluation results obtained in tests using the Process-2000 and the Process-2001 with the same topic set (80 queries). In the CLEF evaluation campaigns the topic set is changed every year and thus direct comparisons of the results of test runs from one year to another cannot be made. By using a large topic set we also hope to get a picture of the robustness of the process to different topics.

The main changes in the process from the year 2000 to the year 2001 are as follows:

- A new process for dictionary look-up and translation of compound words is applied in Process 2001. Compound words are split into components as in Process-2000 but a new grouping strategy for compound components is tested. A more detailed description is presented in 3.1.

---

<sup>2</sup> splitting of compounds was not yet implemented, and non-translated words were handled (using the n-gram method) only in German as a target language.

- A new process for matching proper names and other non-translatable words. An n-gram method (a more detailed description will follow in 3.2) is applied for untranslatable words. The method retrieved six most similar index keys for each untranslatable source language word.
- Stop word lists are applied after the normalization process in Process 2001 contrary to Process-2000 where stop words were eliminated before the morphological analysis. In highly inflectional languages like Finnish the stop word lists would have had to contain all inflected forms if the latter procedure had been applied.
- Normalization of dictionary output added.

## 2.2 Processing of source language words

The UTACLIR system is developed with special regard to compound processing. In many languages (German, Dutch, Finnish and the Scandinavian languages) the components in a multi-word expression are usually written together, not as separate words as in English. In this paper, the term *compound* refers to a multi-word expression where the components are written together. The term *phrase* refers to a case where components are written separately. Therefore the term *compound language* refers to a language where multi-word expressions are compound words rather than phrases, while the term *non-compound language* refers to languages where multi-word expressions are phrases.

The processing of source query keys is based on seven distinct key types.

In the case of non-compound source languages, six key types are recognized.

In case the source key is recognized by morphological software:

- 1) Keys producing only such basic forms which all are stop words. For example the German words *über* (over) is a stop word and thus eliminated before translation.
- 2) Keys producing at least one translatable basic form, for example the German word *Währung* - translated to *currency*.
- 3) Keys producing no translatable basic forms, for example proper names like *Salam, Pierre, Beregovoy, GATT*.

In case the source key is not recognized:

- 4) The key is a stop word. There are very few examples for this key type, in the German topic the word *ex* is such a key that could be a stop word.
- 5) The key is translatable, for example the German words *jordanisch* translated to *Jordanian*.
- 6) The key is untranslatable. For example the German words *Tschetschenien, Bosnien*.

For compound source languages, a seventh key type is added:

- 7) Untranslatable but decomposable compound words. For example the German compound words *Windenergie*, decomposed to *Wind* and *Energie* (English translation: wind energy).

The translation proceeds by utilising the general framework for translating the individual source keys using available linguistic resources. Internally, the recent version of the program uses a three level tree data structure. The first, uppermost level nodes of the tree consist of the original source keys given by the user. The first level also reflects the logical structure of the original source query. The second level nodes contain processed source language strings, for example words generated by morphological programs (like, basic forms or parts of a split compound). Also word analysis information may be saved in the second level nodes. The third and final level of the tree consists of lists of post-processed word-by-word translations (in the target language). Once built, this tree structure can be traversed and interpreted in different ways. The final translated query can be acquired this way and given as the final output of the translation process. Additionally, analysis information (from the second and third level tree nodes) can also be the output.

### ***3 Direct translation results for the whole process***

The main test bed for the UTACLIR system has been the CLEF evaluation forum in the years 2000, 2001 and 2002. The first two years we participated in the bilingual runs using Swedish, Finnish and German topics and the English document collection. In 2002 we changed the source language to English and made bilingual English - Finnish, English - French, English - Dutch and multilingual runs. Each year we made different types of tests. In 2000 compound handling and structuring was tested, in 2001 the revised compound process and the n-gram algorithm. In 2002 we made additional multilingual runs, because the official runs were made using a defective German index.

In addition to these yearly evaluation results that have tested the overall performance of the system, we have wanted to test the robustness and the components of the automated process. For this purpose we have run several tests with a large topic set of 80 topics used in the CLEF evaluation forum in the years 2000 (33 queries) and 2001 (47 queries) in the Finnish, Swedish and German languages.

The average precision values for the two sets of topics (2000 and 2001) are generally better for the 2001 topics. This was also reported by other participants (Peters 2001; Peters et al. 2002).

The effects of the modifications to the process are more significant in the tests with the 2001 topics (Table 1). Also the average precision values for the combined topic sets are better for Process-2001, except for the German language. This exception is due to the fact that the process includes the use of n-grams for all untranslated words. The use of the n-gram algorithm is beneficial in the Swedish and the Finnish processes but not in the German. The average precision values 0.2546 and 0.3476 for the German processes should therefore be compared to the values 0.2639 and 0.3830 where the processes are tested without the use of

the n-gram algorithm. The average precision values were in these cases clearly better. Additional tests with the n-gram algorithm are reported in Section 4.2.

The average precision values for the Swedish process were somewhat irregular. The values for the 2000 topics show a small advantage for Process-2000, but then again there is a significant change to the benefit of Process-2001 for the 2001 topics. Query no. 30 in topics 2000 is troublesome since the relevance assessment gives only 1 relevant document. The average precision value for this individual query changes from 1.000 to 0.0833 with very little changes to the query. If we exclude this query, the Swedish figures for the 2000 topic set would be 0.2307 and 0.2370 instead of 0.2540 / 0.2323.

**Table 1. Average non-interpolated precision for two processes and topic sets**

Topics 2000 N=33	Process-2000	Process-2001	Change %	Monolingual English	Best process % of monolingual
Swe - Eng	<b>0.2540</b>	0.2323	-8.54	0.3880	65.46 %
Fin - Eng	0.2275	<b>0.2469</b>	8.53	0.3880	63.63 %
Ger - Eng	<b>0.2665</b>	0.2546	-4.47	0.3880	68.69 %
Topics 2001 N=47					
Swe - Eng	0.2722	<b>0.3769</b>	38.46	0.4925	76.53 %
Fin - Eng	0.3241	<b>0.3894</b>	20.15	0.4925	79.07 %
Ger - Eng	<b>0.3679</b>	0.3476	-5.52	0.4925	74.70 %
Combined topics N=80					
Swe - Eng	0.2647	<b>0.3172</b>	19.83	0.4494	70.58 %
Fin - Eng	0.2843	<b>0.3306</b>	16.29	0.4494	73.56 %
Ger - Eng	<b>0.3261</b>	0.3092	-5.18	0.4494	72.56 %

Note here that only general-purpose dictionaries were used. Pirkola (1998) has shown that the performance can be further improved by concerted use of general purpose and special purpose dictionaries.

## **4 Source language processing and system components**

In the following sections system components and system features for the handling of source language words are described and evaluated.

### **4.1 Compound words**

Compound source languages require compound handling in CLIR. The dimensions of compound handling include:

- The strategy – whether compounds are (1) kept as such in the process (and omitted if they cannot be translated as such), (2) split into components if they cannot be translated as wholes, or (3) always split and translated by components.
- The treatment of components – whether (1) translated as such or (2) normalised and only then translated.
- Component translation – whether (1) omitted or (2) fuzzy matched if no translation is found.
- Component combination strategy – whether (1) each component is treaded separately, (2) as pairwise combinations or (3) in all combinations derivable from the original compound.
- Target query structuring – whether the component translations form (1) just a large unstructured bag of words, (2) synonym sets based on individual original components, or (3) proximity window based expressions

For compound source languages compound splitting and normalization of compound components are supported. In the handling of compound words, normalized full compound words, if they are not stop words, are first looked up in the dictionary. If a translation, or a set of translations, is available, it is likely to be the best alternative for the source word (a compound or a non-compound). Such compounds are often non-compositional, i.e., the meaning of a compound may be quite different from the meanings of its components (e.g., strawberry). Compound words that do not translate are split into their components and normalized whenever possible. For example the German word *Friedensvertrag* is split into *Frieden* and *Vertrag* the Fuge-element (joining morpheme, see Hedlund et al. 2001) *s* is omitted. Then component translation takes place. The English translation is in this case "peace treaty". If the translation is a phrase, it will be handled as a phrase in the subsequent phases.

The most important changes between Process-2000 and Process-2001 are in the grouping strategy of compounds. We form all consecutive source component pairs and translate these if possible. For example, for a four-component compound a-b-c-d, the component pairs of a-b, b-c, and c-d are formed. Then the pairs are looked up in the dictionary. In the case of several translations, the equivalents are treated as synonyms. All two-component pairs of the translation equivalents are formed for the query. The rationale for this is the left or right branching syntactic structure of compounds. (Hedlund 2002; Malmgren 1994; Warren 1978). In Process-2000 all combinations derivable from the original compound were formed.

The contribution of compound processing in dictionary based cross-language information retrieval is important since around ten percent of content words are compounds in running text after stop word removal in each source language Finnish, Swedish and German (Hedlund 2002).

The following five compound processing modifications were compared. Process 2 was used in CLEF 2001.

Process 1 in Table 2 is a baseline for the tests. In this process no compound splitting or translation of components was performed. Only compounds that could be translated directly using the translation dictionary were used in a similar way as any other single word, that is, a synonym structure for translation alternatives was used, no n-gram matching was used. In the example below of a translated query the German compound *Fussballweltmeisterschaft* (soccer world championship) is not decomposed and thus cannot be translated.

```
#sum( #syn(@fussballweltmeisterschaft fussballweltmeisterschaft) #syn(final end game) #syn(@1994 1994))
```

Process 2 in Table 2 is a test run including the compound processing features. That is, splitting of compounds that do not translate into constituents, normalization of constituents to base forms, the grouping strategy for constituents and translation using a translation dictionary. The target language query (English) gets a structured sub-query using an unordered proximity operator and a window size of 5 + n (n = spaces between words in the phrase). No n-gram matching was used. The compound *Fussballweltmeisterschaft* is decomposed and the components are grouped in pairs before translation. In this case the component pair *Fuss* and *ball* is translated to football, soccer ball, and soccer. The pair *Welt* and *meisterschaft* is translated to world championship. The phrase structure is used in the target language query for "soccer ball", and "world championship". Nonsense combinations like "ball universe" also occur with this pair-wise combination.

```
#sum(#syn(football #uw6(soccer ball) soccer)) #syn(#uw6(ball universe) #uw6(ball everybody) #uw6(ball galaxy) #uw6(ball world) . . .) #uw6(world championship) #syn(final #uw6(end game)) #syn(@1994 1994))
```

Process 3 in Table 2 is similar to process 2 in all other aspects except for the phrase construction in the target language query. Instead of the proximity operator the synonym operator is used. No n-gram matching was used. In the example below instead of the phrases "soccer ball", "ball universe" etc. the translations are synonym sets.

```
#sum(#syn(football #syn(soccer ball soccer) #syn(ball universe ball everybody ball galaxy ball world) . . .) #syn(world championship) #syn(final end game) #syn(@1994 1994))
```

Process 4 in Table 2 includes the best possible translation alternatives for the compounds in the query. This means that ambiguities in the translations provided by the dictionary were manually eliminated. It is, however, disambiguation based solely on the translation of individual compound words in the topic, not on phrases and multi-word concepts or semantic structures in the topic sentences. The example below is an example of the target query in this case.

```
#sum(#syn(football #uw6(soccer ball) soccer) #uw6(world championship) #syn(final #uw6(end game)) #syn(@1994 1994))
```

Process 5 in Table 2 is a monolingual English run. It acts as an upper baseline for the comparison of retrieval performance. Words in the English topic set are normalized in order to match the document index, and stop words are removed. The example below illustrates a query in this case.

#sum(world soccer championship final game world soccer championship 1994)

In Table 2 the average precision results are presented for the five test settings described above. In the rightmost column the best automatic process is compared to the monolingual process.

**Table 2** Effects of compound handling methods using Process-2001

Topics 2000 (N=33)	Process 1. No compound handling process	Process 2. Compound process, phrase structure	Process 3. Compound process, synonym structure	Process 4. Optimal manual compound disambiguation	Process 5. Monolingual English baseline	Best automated process % of monolingual
Swedish	0.2346	0.2242	<b>0.2654</b>	0.2459	0.388	68.40 %
German	0.2079	<b>0.2629</b>	0.2476	0.3282	0.388	67.76 %
Topics 2001 (N=47)						
Swedish	0.3428	0.3465	<b>0.3591</b>	0.3745	0.4925	72.91 %
German	0.3520	0.3830	<b>0.4021</b>	0.3737	0.4925	81.64 %
Combined topics N=80						
Swedish	0.2981	0.2961	<b>0.3204</b>	0.3214	0.4494	71.30 %
German	0.2925	0.3335	<b>0.3384</b>	0.3549	0.4494	75.30 %

The test results indicate that compound handling has an effect on the performance of the retrieval process. Both processes 2 and 3 where we have included a component for compound handling are more effective than process 1 where no compound handling is performed. Particularly the compound handling process where the phrase structure in the target language is substituted by a synonym construction seems to be successful. The phrase structure is discussed in Section 6.2 where target language components are evaluated. The general performance of the compound handling process is rather good.

## 4.2 N-gram techniques for problem word handling

When source language words cannot be translated either due to morphological problems or dictionary deficiencies, their target language equivalents may still often be identified through approximate string matching against the target database index. Among approximate string matching methods, n-grams have proved effective in many applications (Robertson & Willett 1998; Pfeifer, Poersch & Fuhr 1996; Zoebel & Dart, 1995). The following issues are related to the application of n-grams for CLIR:

- Type of n-grams used – the length n, the handling of preceding and trailing spaces, and the formation of n-grams (consecutive vs. inconsecutive characters) may vary
- The source(s) of matching target words may vary – e.g. various target language word lists, the target database index (which may have separate parts for morphologically recognized and unrecognized words).
- The number of matching target language words collected from the word source(s)
- The combination structure of the matching words in the query – e.g. matches for different problem words are inserted individually vs. as synonym groups in the query.

Our tests covered the following cases:

In CLEF 2001 we selected the six highest ranked keys from the result list of n-gram matching for the final queries. For example, the Finnish query 050 contained an untranslatable key *Chiapasissa* (an inflectional word form of the name *Chiapas*). In this case, the six highest ranked keys were *chiapas aphasia shipshape @issaias @aspasia @saipaa*. The first three words are recognized as words by a morphological analyzer and the last three words, marked by '@', unrecognized words. After CLEF 2001 we tested the performance effects of using two, four and six highest ranked n-gram keys in the final queries in Finnish – English CLIR (Pirkola et al. 2003). Of the CLEF 2001 topics (n=50) we selected for this experiment as test queries those topics (n=26) which contained untranslatable keys ( words not found in a translation dictionary or the dictionary of a morphological analyzer). It turned out that the best results were achieved using two highest ranked n-gram keys without using syn-structure (avg. precision 0.441). The use of six highest ranked keys with syn-structure, i.e., the approach we used in CLEF 2001, gave the average precision of 0.417. These findings suggest that the best results are achieved using just a few n-gram keys in the final queries.

In CLEF 2001 we used an n-gram matching technique in which query keys and index terms were split into n-grams consisting both of adjacent characters of the original words as well as *non-adjacent characters separated by one character* in the words. For example, for the word *pharmacology* the following digrams are formed: ph,ha,ar,rm,ma,ac,co,ol,lo,og,gy,pa,hr,am,ra,mc,ao,cl,oo,lg,oy. We call these kinds of n-grams skipgrams. After CLEF 2001, in a research project supporting our CLEF efforts, we devised a novel skipgram matching technique, which we call the *classified skipgram matching technique* (Pirkola et al. 2003). In the technique, digrams are put into categories on the basis of the number of skipped characters. Digrams belonging to the same category are compared to each other but not to digrams belonging to a different category.

We did empirical tests, which showed that the best technique was the classified skipgram technique in which digrams composed of adjacent characters of words formed one category and digrams composed of 1 and 2 skipped characters formed another category. In the empirical tests, this technique was compared with

the conventional n-gram technique using adjacent characters as n-grams, as well as the skipgram technique we used in CLEF 2001 (as explained above). Several types of words and word pairs were studied. English, German, and Swedish query keys were matched against their Finnish spelling variants in a target word list of 119,000 Finnish words. In all tests done, the classified skipgram technique outperformed the conventional n-gram matching technique as well as the skipgram technique applied in CLEF 2001. Based on these results we have planned to incorporate the classified skipgram technique into UTACLIR.

The highest precision improvements were achieved when the start and end spaces were used as constituent characters of n-grams. For classified skipgrams the relative improvement percentages with respect to conventional n-grams were up to 49.7% (English-to-Finnish matching), 20.7% (German-to-Finnish matching), and 17.1% (Swedish-to-Finnish matching). The results were statistically significant at the levels of 0.01-0.001.

## **5 Translation resources**

Translation dictionaries greatly affect the performance of dictionary-based CLIR systems. Dimensions related to their usability and effects include:

- The number of headwords affects the frequency of ex-dictionary problem words
- The types of headwords – whether mainly individual words vs. also compounds and phrases are provided as headwords.
- The number and order of translation equivalents – exhaustivity tends to bring rare and thus hardly useful senses into CLIR queries; ordering by sense frequency could support a translation strategy that is less comprehensive.
- Structure – how easy is it to separate translations from all other content like etymology and examples? This affects the simplicity and accuracy of translation.

The translation resources used in the tests are machine-readable bilingual translation dictionaries and translation tables.

- Motcom Swedish - English translation dictionary (60,000 entries) by Kielikone plc. Finland. (<http://www.kielikone.fi/english/>)
- Motcom Finnish – English - Finnish translation dictionary (110,000 entries) by Kielikone plc. Finland
- Oxford Duden German - English translation dictionary (260,000 entries) Translation tables were constructed for the automated process.
- Motcom GlobalDix multilingual translation dictionary (18 languages, total number of words 665,000) by Kielikone plc. Finland

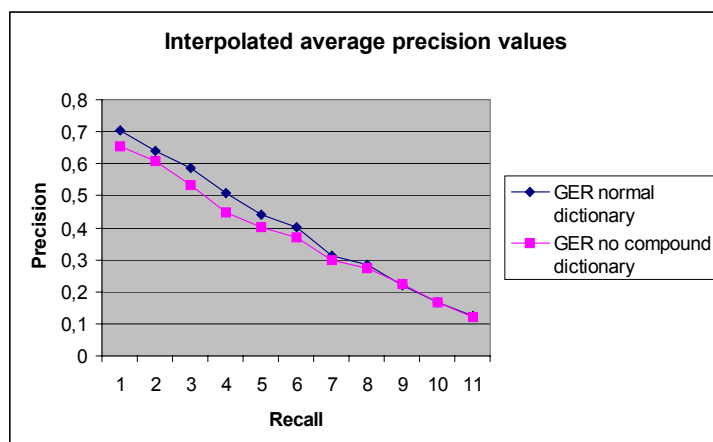
- Motcom Swedish – Finnish bilingual translation dictionary (80,000 entries) by Kielikone plc. Finland
- Motcom Finnish – Swedish bilingual translation dictionary (55,000 entries) by Kielikone plc. Finland
- Motcom German – Finnish bilingual translation dictionary (85,000 entries) by Kielikone plc. Finland
- Norstedts Tyska Ordbok: German – Swedish – German translation dictionary (127,000 words and phrases) by Norstedts Ordbok AB, Sweden

*The effect of removing compounds from the dictionary – simulating a small dictionary.* For German – English we have tested two types of dictionaries (two runs). Using the Duden German-English dictionary (260,000 words), two translation tables for 50 CLEF topics (title and description field) were created. The first included all translations from the dictionary. The second translation table contained the same data, except that all direct translations of compounds were excluded. Altogether 64 individual compounds were removed. 11 topics did not contain any compounds. For the remaining 39 topics the average number of individual compounds is 1,6 per topic.

The test with two dictionaries for the German to English runs (see Table 3 and Figure 2) shows that the UTACLIR process also works well with a limited dictionary. On the other hand, the advantage of a direct translation of compounds is inevitable. Our method for handling compounds works as a good and necessary complement, since no dictionary, not even a comprehensive one, holds entries for all compounds. Compound splitting was necessary also in several queries when the comprehensive dictionary was used. The queries based on the short dictionary became very long since all compounds were split into their components. When all alternative translations for the components are combined in a phrase in the target language query, the number of combinations may be high. Nonsense combinations also occurred quite frequently. For example, for the German word *Schildkröte* (tortoise), we get combinations like "shield toad", "shield creature", "nameplate toad", "badge creature" etc. On the other hand, generally, the process can be said to work as expected because of the existing relevant combinations.

**Table 3. Average precision for results for German-English bilingual runs using a comprehensive and a short dictionary.**

<b>Testrun using Process-2001</b>	<b>Average precision</b>
German-English with a comprehensive dictionary	0.3474
German-English with a short dictionary (compounds eliminated)	0.3054



**Figure 2 Interpolated average precision values for German-English bilingual runs using a comprehensive and a short dictionary.**

*The effect of dictionary size.* Changing the dictionary to GlobalDix in Process-2002 affects the performance of the bilingual runs. It is impossible to test this effect without a parallel dictionary. We have a parallel English - Finnish translation dictionary, MOTCOM bilingual English – Finnish dictionary with 110,000 entries, and thus it is possible to compare the effect of the dictionary. The MOTCOM GlobalDix dictionary is much more constrained than the bilingual one. GlobalDix has only 26,000 English – Finnish entries.

The baseline is the run with the GlobalDix dictionary. The average precision was 0.246 with the GlobalDix dictionary, and 0.326 with the bilingual dictionary (Table 4). The result was 32.5 % better with the bilingual dictionary than with GlobalDix. We can conclude that the effect of the translation dictionary on the result is great, and GlobalDix is not an optimal dictionary for English – Finnish runs. We did not test the exclusion of compounds in this setting.

**Table 4. Average precision results for English - Finnish bilingual runs using alternative resources**

	Average precision	Difference	Difference %
GlobalDix dictionary	0.246		
MOT bilingual dictionary	0.326	+0.080	+32.5%

## 6 Target language components

The target language query formulation supports normalization, stemming, synonym structuring of translation equivalents and phrase-based structuring of target phrases. The process is described in detail in Hedlund et al. (2001b; 2002b) and Hedlund (2002). The Pirkola-method used for structuring of target language queries is described in Pirkola (1998), and the n-gram based matching of untranslatable words in Pirkola et al. (2003).

The English, Finnish, German and Swedish databases were indexed using morphological analyzers (ENGTWOL, FINTWOL, GERTWOL and SWETWOL by Lingsoft plc.) which produce three basic cases. First, the input word was recognized by the analyzer and (one) basic form was produced in the index. Second, for homographic word forms (e.g. English words saw, left) sometimes more than one index key was produced ( saw / see, leave/ left respectively). The first two cases form the database index of recognized words. The third case consists of input words not recognized by the analyzer (e.g. basetsane). Such words were indexed as such, preceded by a special symbol (@basetsane), thus constituting the index of unrecognized words.

### 6.1 Structuring of queries

To be able to test the effect of query structuring we have constructed both structured and unstructured queries for all language pairs. Query structure is the syntactic structure of a query expression, as expressed by query operators and parentheses. In this study, queries with a single operand and no differentiated relations between search keys are called unstructured queries, and queries with the synonym operator combining search keys translated from the same source language word are called structured queries. The synonym operator syntax is: #syn( $T_1 \dots T_n$ ) where  $T_i (1 \leq i \leq n)$  are terms. The terms within this operator are treated as instances of the same term for InQuery's belief score computation. For example, the translation of the Swedish word *möte* becomes #syn(meeting encounter crossing appointment date).

**Table 5. Average non-interpolated precision values of structured and unstructured queries**

Topic set 2000 N=33	Unstructured queries no n-gram	Structured queries no n-gram	Change in %
Swe - Eng	0.2015	<b>0.2242</b>	11.3 %
Fin - Eng	0.1609	<b>0.2150</b>	33.6 %
Ger - Eng	0.2097	<b>0.2639</b>	25.8 %
Topic set 2001 N=47			
Swe - Eng	<b>0.3466</b>	0.3465	0.0 %
Fin - Eng	0.2407	<b>0.3443</b>	43.0 %
Ger - Eng	0.3242	<b>0.3830</b>	18.1 %
Combined N=80			
Swe - Eng	0.2867	<b>0.2961</b>	3.3 %
Fin - Eng	0.2078	<b>0.2910</b>	40.0 %
Ger - Eng	0.2770	<b>0.3335</b>	20.4 %

The results in Table 5 indicate that structured queries perform better. The difference in performance is greatest for the language pair Finnish-English. The Finnish structured queries compared to the unstructured ones show a better performance of 0.0832 in average precision. The German structured queries show a better average precision result of 0.0565, and for Swedish the change is 0.0094.

## 6.2 Compounds treated as phrases

Compound translations often form phrases in the target language English. This was taken into account in Process-2000 by using a proximity operator with quite a narrow window size. Process-2001 was modified by a more flexible proximity operator and a slightly broader window size. That is, the proximity operator was changed from OD (ordered window) to UW (unordered window) which allows for free word order in the target phrases. The window size was set to 5 + n, where n = the number of spaces between words in the phrase. The effect of the phrase structure is presented in Section 4.1, Table 2. Since phrase-based structuring concerns compounds, it was evaluated as an additional feature together with the compound handling component.

Originally we hypothesized that allowing phrase structure in target language queries would yield the best performance. However, the findings in Section 4.1 (process 3, Table 2) indicate that when the phrase structure in the translated target language query was substituted by a synonym structure the results were beneficial for the latter query setting. The following examples from the individual query 069 illustrate this.

```
Query 069 #sum(clone #syn(ethics #uw6(ethical work) ) #syn(practical #uw6(general practitioner) useful
#uw6(in practice) #uw6(practical mind) practical virtual) #syn(use employment take application)
#syn(@klonens klonens) #syn(ethical ethical) argument )) ;
```

result: average precision 0,0649 for phrase structure

Query 069 #sum(clone #syn(ethics #syn (ethical work) ) #syn(practical #syn (general practitioner) useful #syn (in practice) #syn (practical mind) practical virtual) #syn(use employment take application) #syn(@klonens klonens) #syn(ethical ethical) argument ));

result: average precision 0,3036 for synonym structure

Similar results have been given in Pirkola et al. (2003). The results are consistent with what has been reported in earlier studies on English monolingual retrieval. For example, Mitra et al. (1997) found that phrase-based searching might decrease retrieval performance. It should be noted that this monolingual component is involved in CLIR. However, one should also note that the situation is more complicated in CLIR, as phrase structure also has a clear disambiguation effect (Ballesteros & Croft 1997).

### **7 Multilingual results with the new UTACLIR system**

There are several different strategies for merging the results obtained from distinct databases. The simplest of them is *the Round Robin approach*, which means that one document of every result list is taken, one by one from each, until there are as many documents as needed. This is based on the assumption that the distribution of relevant documents in the lists is not known, because the scores are not comparable, and there is no way to compare them. *The raw score approach* assumes that document scores are comparable across separate collections. *The rank based approach* depends on the fact that the relationship between probability of relevance and the log of the rank of a document can be approximated by a linear function. Merging can subsequently be based on the re-estimated probability of relevance. The actual score can then be applied only to ranked documents, but the merging is based on the rank, not on the score. (Hiemstra & al. 2001)

The multilingual runs were performed by merging the result sets of bilingual English – French, English – German, English – Italian, English – Spanish runs, and a monolingual English run. The average precision of the bilingual runs varied between 0.201 and 0.246. In our first official CLEF 2002 multilingual run we applied the raw score approach. The average precision of this run was 0.1637. The result using the Round Robin merging approach was 0.1166.

We made additional multilingual runs, because the German index was defective in the official CLEF 2002 runs. The average precision of the additional multilingual run with the raw score approach was 0.1829, and with the Round Robin approach 0.1612.

The results of the multilingual runs are poor compared to the results of the runs from which they have been merged. Both of the merging strategies proved unsuccessful in the multilingual CLEF run. Besides developing the UTACLIR translation process, we need to investigate more advanced merging strategies.

## **8 Transitive translation results**

Motcom dictionaries (Finnish-Swedish and Swedish-Finnish). In translating from German into Swedish Besides the bilingual translation processes, we have experimented with transitive processes, i.e., processes using an intermediate (or pivot) language between the source and the target languages. As source language we used Finnish, Swedish and German, as pivot language Finnish and Swedish, and as target language English. The transitive processes were constructed on the basis of the bilingual 2001-processes. The same resources as in the bilingual processes were used at each step. The translation dictionaries were, a bilingual wordlist compiled from Norstedts German-Swedish dictionary was used.

The same approaches as in the bilingual processes were used in the transitive processes. There were some modifications, however. For example for compounds that were not translated as full compounds, we did not use the proximity operator UW to combine the translation equivalents of the components. These were only combined by the synonym operator. In other words, the German source word ‘Weltwetter’ (‘world weather’) was translated into English via Finnish as follows: #syn(earth world globe weather better) (c.f. the bilingual process: #syn(#uw6(earth weather) #uw6(earth better) #uw6(world weather) #uw6(world better) #uw6(globe weather) #uw6(globe better)) ).

The bilingual runs Swedish-English, Finnish-English and German-English were used as the baseline. The results of the transitive runs were compared to those of the bilingual runs to find out how much is lost in effectiveness when translation is performed through an intermediate language.

The results of the transitive and the baseline bilingual runs, measured in average precision, are presented in Table 6. The transitive runs, on the average, performed well compared to the bilingual ones. There was, however, great variance among the results of individual topics. For some topics, bilingual and transitive queries performed equally, for some the bilingual queries performed better, and for some the transitive queries gave better results. In the 2001 Swedish topics the greatest difference between the runs was 99.4 % (0.006 for the bilingual, 1.000 for the transitive).

In our experiments, transitive translations, in comparison to bilingual ones, performed better than in previous studies (Ballesteros 2000; Gollins & Sanderson 2001a; 2001b). It should be noted that the results were achieved by the basic translation process, using only query structuring - no other measures, e.g. triangulation (Gollins & Sanderson 2001a), were needed.

**Table 6 Average precision for bilingual and transitive translation for year 2000 and 2001 Swedish, Finnish and German topics**

TRANSLATION TYPE	2000 TOPICS (N=33)			2001 TOPICS (N=47)		
	Average precision	Diff	Diff %	Average precision	Diff	Diff %
ENG MONOL	0.361			0.480		
SWE-ENG	0.236			0.373		
SWE-FIN-ENG	0.232	-0.004	-1.7 %	0.402	+0.029	+7.8 %
FIN-ENG	0.260			0.405		
FIN-SWE-ENG	0.204	-0.056	-21.6 %	0.359	-0.046	-11.4 %
GER-ENG	0.235			0.389		
GER-FIN-ENG	0.175	-0.060	-25.5 %	0.323	-0.066	-17.0 %
GER-SWE-ENG	0.195	-0.040	-17.0 %	0.326	-0.063	-16.2 %

Dictionaries are an essential component of the transitive translation process. The performance of a query may crucially depend on whether a certain source or pivot word is translated, or if it is translated in a certain way. A word might not be translated simply because it is not included in the dictionary. Another possibility is that the dictionary has the word as an entry but it is not found since it is not in the form searched for (e.g. singular form vs. plural form, basic form vs. inflected form). E.g. in the process Finnish-Swedish-English, after the first translation phase and the word form normalization, the Swedish word 'medelhav' (Mediterranean) was unsuccessfully matched against the Swedish-English dictionary entry 'medelhavet' (definite form of a noun).

Query words may also be translated deficiently, e.g., some essential translation equivalents may be missing. For a topic on Persian Gulf war syndrome, it was, e.g., quite essential that the word *gulf* was in the final target query. In the process Finnish-Swedish-English, the Swedish word 'vik' (gulf) was given the translation equivalents *bay* and *creek* (by the Swedish-English dictionary). Average precision of the English target query was 0.204. The value would have been 0.621 (not much below that of the corresponding bilingual translation: 0.689) if the word *gulf* had not been neglected by the dictionary.

Incompatibility is a factor contributing negatively to the performance of transitive translation. It may be manifested between dictionary entries and word forms produced by the morphological process, as above,

but also elsewhere in the process. For example, the translation equivalent produced by the process and the word used in the relevant documents may be incompatible. The word given by the dictionary may in itself be correct but some other word is favoured in relevant documents. During our experiments, cases of incompatibility were also found between words used in the source language and words used in the target language. In many cases, the source language (Finnish) preferred nouns when the target language (English) preferred adjectives in the same expression. E.g., the Finnish expression for *European Economic Area* is *Euroopan Talousalue* (Europe Economy Area). Average precision of the transitively translated Finnish topic (CLEF 021) containing the original expression was only 0.029. When the nouns of the expression were changed into the corresponding Finnish adjectives in the topic, the value was raised to 0.151.

## 9 Discussion and Conclusion

The contributions of this study are the following:

- Performance analysis results of a dictionary-based CLIR system for several language pairs and a large topic set
- An analysis of the contribution of individual components of the process: the effectiveness of compound handling, proper name matching, query structuring and translation dictionaries
- Performance analysis results of multilingual tests
- Performance analysis results of a transitive translation process

In general, the test results indicate that the UTACLIR process is quite robust and applicable in the sense that several languages are used with relatively small differences in performance. However, there is a difference in performance regarding the two topic sets used. The average precision values for the queries in the topic set 2000 (topics 1-33) are lower in each of the runs and regardless of topic language. This is also true for the monolingual run. There is a drop in average precision of the cross-language queries with respect to the monolingual baseline, which was expected.

By testing the individual component performance, we could collect valuable information for the further development of the UTACLIR system. The tests with structuring by the *Pirkola method* of the queries indicate that structuring is a good way to reduce the effect of ambiguity caused by several dictionary translation equivalents for a source language word. This is true for all the source languages, but it is particularly noticeable for Finnish and German where the translation dictionaries are comprehensive and contain a lot of translation equivalents.

The compound handling process for compound rich languages is important. Splitting compounds into constituents enables a translation of the constituents, which often are content bearing words. In this study where the target language is English, a setting allowing a phrase structure in the translated target language

query was thought to be the best solution. However, the findings indicate that when the phrase structure in the translated target language query was substituted by combining the translations of the split compound components in a synonym structure, the results were beneficial.

The n-gram algorithm was implemented in the process in order to handle untranslated words, some of them most likely proper names. The process was particularly successful for Finnish where proper names usually appear in inflected form and where matching to the target document index is therefore difficult. However, additional noise was added to the queries by the n-gram algorithm, which resulted in a decrease in performance for German. The n-gram algorithm in this setting returned a set of six most similar words - three most similar from the index of morphologically recognized words and three from the index of morphologically unrecognized words. This choice was probably not an ideal one.

The findings in this study for the bilingual processes indicate that a similar process may also be applied in a multilingual environment. The basic structure of the process is applicable to other language pairs, and the specific language dependent features in the component processes can be adapted for each language pair. Result merging is a significant step in multilingual run, and we must also pay attention to it.

The following two conclusions can be drawn from our transitive query translation experiments: 1) it is possible to achieve good effectiveness in transitive translation, and 2) good effectiveness can be achieved using fairly simple methods. In developing and analyzing the processes, we found factors contributing harmfully to the performance of transitive translation. Though such factors are also common in bilingual query translation, they are more likely to appear here because of the additional translation phases needed. They may concern the translation process itself or its components, e.g. dictionaries used. Defects in the process may be noticeable, yet tolerated, or hidden, hindering the process from working as it is supposed to. The former are also potential points for future improvements of the process. Problems connected with the process itself can probably be solved - if only they are discovered. Problems with dictionaries may be more difficult to solve. In many cases, the only solution would be to change the current dictionary for a better one, which often is only a theoretical option.

As a general conclusion, the results in this series of studies also indicate a positive outcome for the combination of linguistic tools and a matching technique based on approximate string matching and structuring of queries. The linguistic approaches using normalization, and compound handling as well as all the other components have a considerable value in improving effectiveness, but the optimization of the combined effects is a challenge and requires further research.

Finally, the UTACLIR processes show that performance comparable to that of other approaches can be achieved by simple translation techniques. After all, the UTACLIR process is a single pass translation and construction process - without (pseudo) relevance feedback, query biased summaries, or machine

translation. By applying such techniques, the results may be further improved.

## References

- Airio E, Keskustalo H, Hedlund T and Pirkola A. (2002) UTACLIR @ CLEF 2002: Towards a Unified Translation Process Model. In Peters C, ed. Working Notes for the CLEF 2002 Workshop, 19-20 September, Rome, Italy 2002, pp. 51-58. <http://clef.iei.pi.cnr.it/> (accessed March 8<sup>th</sup>, 2003)
- Ballesteros L (2000) Cross language retrieval via transitive translation. In Croft W B ed. Advances in information retrieval: Recent research from the CIIR, pp. 203-234. Boston: Kluwer Academic Publishers.
- Ballesteros L and Croft B (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20<sup>th</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, July 27-31, 1997, pp. 84-91.
- Davis M and Ogden W C (1997) QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System. In Proceedings of the 20<sup>th</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, July 27-31, 1997, pp. 92-98.
- Gachot D, Lange E and Yang J (1998) In Grefenstette G ed. Cross-Language Information Retrieval, pp. 105-118. Boston: Kluwer Academic Publishers.
- Gollins T and Sanderson M (2001a) Improving Cross Language Information Retrieval with Triangulated Translation. In Proceedings of the 24<sup>th</sup> ACM/SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, September 9-13, 2001, pp. 90-95.
- Gollins T and Sanderson M (2001b) Sheffield University CLEF 2000 submission - bilingual track: German to English. In Peters C ed. Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069, Springer-Verlag, Berlin 2001, pp. 245-252.
- Hedlund T (2002) Compounds in dictionary-based cross-language information retrieval. Information Research, 7(2) <http://InformationR.net/ir/7-2/paper128.html>. (accessed March 8<sup>th</sup>, 2003).
- Hedlund T, Keskustalo H, Airio E and Pirkola A (2002a) UTACLIR - an Extendable Query Translation System. Paper presented at the ACM SIGIR Workshop for Cross-Language Information Retrieval, August 15<sup>th</sup> 2002 in Tampere, Finland
- Hedlund T, Keskustalo H, Pirkola A., Airio E, and Järvelin K (2002b) UTACLIR @ CLEF 2001 - Effects of compound splitting and n-gram techniques. In Peters C, Braschler M, Gonzalo J and Kluck M eds. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Lecture Notes in Computer Science 2406, Springer-Verlag, Berlin, 2002. pp. 118-136.
- Hedlund T, Pirkola A. and Järvelin K (2001a) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information Processing & Management vol.37(1) pp.147-161.
- Hedlund T, Keskustalo H, Pirkola A, Sepponen M and Järvelin K (2001b) Bilingual tests with Swedish, Finnish and German queries: Dealing with morphology, compound words and query structuring. In Peters C ed. Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069, Springer-Verlag, Berlin, 2001. pp. 211-225.

- Hiemstra D, Kraaij W, Pohlmann R and Westerveld T (2001) Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In Peters C ed. *Cross-language information retrieval and evaluation: Proceedings of the CLEF 2000 Workshop*, Lecture Notes in Computer Science 2069. Springer-Verlag, Berlin, 2001. pp. 102-115.
- Hull D and Grefenstette G (1996) Querying across languages: A dictionary-based approach to multilingual information retrieval. In: *Proceedings of the 19<sup>th</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland, August 18-22. pp. 49-57.
- Keskustalo H, Hedlund T and Airio E (2002) UTACLIR - General query translation framework for several language pairs. Demoposter, in: *Proceedings of the 25<sup>th</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, August 11-15<sup>th</sup>, 2002. pp. 448. Demonstration.
- Malmgren S G (1994) *Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser.* [Swedish lexicology. Words, word formation, dictionaries and word databases.] Lund: Studentlitteratur
- Mitra M, Buckley C, Singhal A and Cardie C (1997) An analysis of statistical and syntactic phrases. In *Proceedings of RIAO'97, Computer Assisted Information Searching on the Internet*, Montreal, Canada, 1997. pp. 200-214.
- Nie J-Y and Jin F (2002) Merging Different Languages in a Single Document Collection. In Peters C ed. *Working Notes for the CLEF 2002 Workshop*, September 19-20<sup>th</sup>, Rome, Italy 2002. pp. 59-62. <http://clef.iei.pi.cnr.it/> (accessed March 8<sup>th</sup>, 2003)
- Oard D and Diekema A (1998) Cross-Language Information Retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223-256.
- Pfeifer U, Poersch T and Fuhr N (1996) Retrieval effectiveness of proper name search methods. *Information Processing & Management*, 32:667-679.
- Peters C ed. (2001) *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop*, Lecture Notes in Computer Science 2069, Springer-Verlag, Berlin 2001.
- Peters C, Braschler M, Gonzalo J, Kluck M. eds. (2002) *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001.* Lecture Notes in Computer Science 2406, Springer-Verlag, Berlin, 2002.
- Pirkola A (1998) The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21<sup>st</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 24-28<sup>th</sup> 1998. pp. 55-63.
- Pirkola A (1999) *Studies on linguistic problems and methods in text retrieval.* Ph.D. Thesis, University of Tampere. *Acta Universitatis Tamperensis* 672, 1999.
- Pirkola A, Hedlund T, Keskustalo H and Järvelin K (2001) Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information Retrieval*, 4(3/4):209-230.
- Pirkola A, Puolamäki D and Järvelin K (2003) Applying query structuring in cross-language retrieval. *Information Processing & Management*, 39:391-402.
- Robertson A M and Willett P (1998) Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48-69.

Sheridan P, Ballerini J P and Schäuble P (1998) Building a large multilingual test collection from comparable news documents. In Grefenstette G ed. *Cross-Language Information Retrieval*, Boston, Kluwer Academic Publishers 1998. pp. 137-150.

Sperer R and Oard DW (2000). Structure translation for cross-language IR. In *Proceedings of the 23<sup>rd</sup> Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, July 24-28, 2000 pp.120-127.

Warren B (1978) *Semantic patterns of noun-noun compounds*. Göteborg: Acta Universitatis Gothoburgensis. (Gothenburg studies in English 41).

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