TURID HEDLUND

Dictionary-Based Cross-Language Information Retrieval

Principles, System Design and Evaluation

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Sciences of the University of Tampere,
for public discussion in the Auditorium Pinni B 1096,
Kanslerinrinne 1, Tampere, on November 8th, 2003, at 12 o'clock.

Acta Universitatis Tamperensis 962
University of Tampere
Tampere 2003

ACADEMIC DISSERTATION
University of Tampere, Department of Information Studies
Finland

Distribution

University of Tampere
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 215 6055
Fax +358 3 215 7685
taju@uta.fi
http://granum.uta.fi

Cover design by
Juha Siro

Printed dissertation
Acta Universitatis Tamperensis 962
ISBN 951-44-5789-7
ISSN 1455-1616

Electronic dissertation
Acta Electronica Universitatis Tamperensis 288
ISBN 951-44-5790-0
ISSN 1456-954X
http://acta.uta.fi

Tampereen yliopistopaino Oy Juvenes Print
Tampere 2003

Acknowledgements

This thesis would not have been written without functioning networks and the possibilities to communicate and work from distance that the Internet provides. Most of my research work was done in my home in Grankulla and the communication with the research group and my supervisor at the Department of Information Studies at the University of Tampere was done using e-mail. The distance also made the physical meetings, the monthly seminars of the FIRE research group and the discussions with my supervisor very intensive and important days. Even though I did not spend much time physically working at the department, my impressions of the research environment in Tampere are only positive.

However, a good, functioning research environment is nothing without the people working in it. I had the privilege to have as my supervisor Professor Kalervo Järvelin, a truly creative and inspiring person and a wizard in proposing solutions to research problems. I am deeply grateful to my co-authors Ari Pirkola, Heikki Keskustalo, Eija Airio and Kalervo Järvelin. To work with them in the CLEF campaign and see the UTACLIR translation system take form as a computer program was a fight to keep the deadlines, but also shared moments of joy when the results were positive. I also want to thank all the researchers in the FIRE research group at the Department of Information Studies, who at numerous occasions made valuable comments to every part and article in the thesis. A special thanks goes to Raija Lehtokangas for her comments to the final thesis and to Bemmu Sepponen who wrote the first program code for the Swedish-English translation.

During my work with the thesis I have had the pleasure to discuss many philosophical aspects of research work with Dr. Leif Andersson from the University of Helsinki. He also for the first time introduced me to research literature on information retrieval. I remember Carol Peters, the main organiser of the CLEF evaluation forum, with gratitude for her inspiring work and positive attitude.

I am also very grateful to the library director Maria Schröder and the colleagues at the library of the Swedish School of Economics and Business Administration, who have been supportive and patient through the years. I would also like to thank professor Bo-Christer Björk and the members of the SciX research group at the Department of Management and Organisation.

Last but not least I would like to remember and keep in my hearth my family - Torolf, Linda, Johanna and Fredrik. Thank you!

Grankulla 28.9.2003

Turid Hedlund

Abstract

The research problems of the thesis relate to the Scandinavian language Swedish. When the research work on this thesis started, there was very limited knowledge on information retrieval or cross-language information retrieval research in Swedish. The linguistic features of this and other compound rich languages indicate that research focusing on languages of other types than English is of great importance. One problem was also the lack of automated dictionary-based systems for query translation of Scandinavian languages and other compound rich languages.

Firstly, cross-language information retrieval problems for non-English languages, particularly Swedish are discussed. In the article the need to extend research on information retrieval techniques to undertreated languages is demonstrated.

Secondly, one of the main problems identified for Swedish, the frequent presence of compounds is discussed in detail and solutions are proposed. Retrieval efficiency may be improved by splitting not directly translatable compounds into constituents using morphological analysis programs and by normalising the constituents into base form before translation using machine-readable dictionaries. This solution is tested for 80 cross-language information retrieval queries.

Thirdly, this thesis deals with bilingual natural language information retrieval techniques where English is the target or document language and Swedish, Finnish and German are source or query languages. The system design of the UTACLIR, an extendable bilingual dictionary-based query translation system, is presented. The approach is to apply linguistic tools in an automated dictionary-based system able to handle several languages.

Fourthly, the performance of the system is evaluated in international evaluation campaigns and shown effective. The automated CLIR process is also tested for the performance of its components. The tests with structuring 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 is particularly notable for Finnish and German where the translation dictionaries used in the study were comprehensive. Compound handling for the compound rich source languages Swedish, German and Finnish is found beneficial to the system performance. An n-gram based algorithm was implemented in the process in order to solve the problem of untranslatable words, such as proper names. The process was particularly successful for the Finnish language where proper names usually appear in inflected forms and where matching to the target language document index therefore is difficult.

TABLE OF CONTENTS

Original research papers 9

1 Introduction..11

PART I 14

2 Linguistic features in information retrieval 15

2.1 Inflectional and derivational morphology..15

2.2 Compounds and phrases...16

2.3 Lexical meaning and ambiguity 18

3 Natural language tools for information retrieval...........................20

3.1 Stemmers..20

3.2 Normalisers 22

3.3 Handling of joining morphemes...22

3.4 Stop words..23

3.5 Other forms of natural language analysis in information retrieval 23

4 Cross-language information retrieval - problems and approaches.............25

4.1. Dictionary-based methods 27

4.2 Research results in cross-language information retrieval 31

5 The UTACLIR query translation system 34

5.1 The general UTACLIR framework for query construction............34

5.2 General description of the UTACLIR system...................................35

6 The InQuery Information Retrieval System 39

7 Evaluation methods and test environment for cross-language information retrieval 42

7.1 Evaluation measures 42

7.2 Topic creation and evaluation for cross-language information retrieval 44

8 Summary of the studies 46

8.1 Summary of Part II 46

8.2 Summary of Part III...50

9 Discussion and conclusion 56

References..59

Appendices..65

CLEF 2000, English topics..66

CLEF 2001 English topics..76

PART II

PART III

Original research papers

PART II

Hedlund, T., Pirkola, A., & Järvelin, K. (2001). 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. 2002. Compounds in dictionary-based cross-language iinformation retrieval, *Information Research*, 7(2).
http://InformationR.net/ir/7-2/paper128.html

PART III

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 structuring. In Peters, C. Ed. *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Revised Papers.* Lecture Notes in Computer Science 2069, Berlin: Springer, 2001. pp. 211-225.

Hedlund T., Keskustalo H., Pirkola A., Airio E., & Järvelin K. (2002). 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, Berlin: Springer, 2002. pp. 118-136.

Hedlund,T., Pirkola, A, Keskustalo, H., Airio, E. & Järvelin K. (2002). Cross-language information retrieval: Using multiple language pairs. In Bothma T., Kaniki A. Eds. *Progress in Library and Information Science in Southern Africa. Proceedings of the second biennial DISSAnet Conference.* 24-25 October 2002, Farm Inn, Pretoria, South Africa.

1 Introduction

Information retrieval as a concept is linked to the user and the user's information need. An information need is specified in information retrieval as a *request*. The information content retrieved by an information retrieval system should be relevant to the request. Retrieval systems are able to retrieve information in several formats, including bibliographical information in the form of references, text documents, as well as images and spoken documents. This is in accordance with the development in information technology. Information is available to us due to network techniques in a variety of formats and in an ever-increasing volume.

Information retrieval in this study focuses on text retrieval - text documents. Text retrieval can also be seen as closely related to the larger field of linguistics and natural language processing (Strzalkowski 1999). Pirkola (1999) gives in his study a thorough review of general problems and methods in text retrieval especially from the linguistic point of view. He states that it is the task of information retrieval research to recognise the problems in information retrieval related to language and to find solutions to them. Complementary to this goal, statistical techniques still have a leading role in information retrieval research and applications. Retrieval models like the vector space and probabilistic models were developed as early as in the sixties and seventies (Salton, Wong & Yang 1975; Salton & McGill 1983; Maron & Kuhns 1960; Robertson 1977). The development of computer techniques has enabled the earlier theories to be applied and tested in environments of realistic size and also in operational systems.

Cross-language information retrieval is defined as the retrieval of documents in another language than the language of the request. The language of the request is the *source language* and the language of the documents is the *target language*.

International communication and the multitude of information in several languages require information retrieval systems that can cross language borders. Internet is by far the best example of an environment where mediated access to network resources in different languages is needed. Many people have a reading ability in a language different from his/her native language, but writing and query formulation can be more difficult. Even in the case of very close languages, e.g., between Scandinavian languages (Swedish, Norwegian and Danish), where understanding and reading cause no problem, formulating queries to an information retrieval system can be very difficult. For Gerald Salton (1970), who reported one of the earliest experimental results for cross-language text retrieval, the mass of information resources was probably not the main problem. However, it exemplifies that information retrieval research very early tried to solve the problem of multilingual texts.

The expressions multilingual, cross-language, cross-lingual and cross-linguistic have been used with slightly different definitions. The first workshop on cross-language retrieval systems was held at the ACM SIGIR Conference on Research and Development in Information Retrieval in 1996. This initiative and the desire to build better cross-language systems resulted in 1997 in a cross-language track at the Sixth Text Retrieval Conference, TREC-6 (Harman et al. 2001). The terminology was also clarified as the term cross-language information retrieval was agreed on as the best single description (Oard 1997).

A multilingual collection contains documents in different languages or even individual documents that contain text in more than one language. The concept Cross-language retrieval clearly distinguishes retrieval crossing language borders from monolingual information retrieval. *Cross-language information retrieval* specifically deals with the problem of presenting an information retrieval task in one language and retrieving documents in one or more 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, or there are multiple monolingual target collections in different languages and requests are expressed in a language different from the collection.

The research problems of the thesis relate to the introduction of a new language and language family in information retrieval and cross-language information retrieval research, the Scandinavian language Swedish. The linguistic features of this and other compound rich languages indicate that research focusing on languages of other types than English is of great importance. The problem was also the lack of automated dictionary-based systems for query translation of Scandinavian languages and other compound rich languages. Language features like morphology, semantics and compound words, among others, have to be taken into account when developing systems for information retrieval and cross-language information retrieval.

The objective of this thesis is to contribute to the area of cross-language information retrieval, firstly, by developing and evaluating a new robust query translation system for cross-language information retrieval, UTACLIR. Secondly, the thesis contributes by focusing on features in languages that are important in dictionary based cross-language information retrieval. The approach is to apply linguistic tools in an automated dictionary-based system able to handle several languages. Thirdly, the research in this thesis also contributes by introducing a new language and language family in cross-language and information retrieval research, the Scandinavian language Swedish.

In this thesis a mix of linguistic and statistical techniques is employed in the development of a system for dictionary-based cross-language information retrieval. The results have been evaluated on an international research forum, the Cross-Language Evaluation Forum, CLEF 2000, 2001.

The thesis consists of three parts. The first part is an introduction and summary to the following two empirical parts. The second part is linguistically oriented, discussing relevant aspects in Swedish from cross-language information retrieval point of view. It also discusses compound words as features in compound-rich languages, and the way to handle source language compounds in bilingual cross-language information retrieval. The third part is system-oriented, and the objectives are to develop and evaluate a new system for automated cross-language retrieval. Linguistic features of several morphologically rich languages are taken into account in the development phase. The performance of the system and its different components are tested. It is important not to focus solely on the overall performance of the system, but also to be able to evaluate individual system components in order to learn about them and their interaction.

PART I

The first part of this thesis consists of Sections 2 to 9. They are intended as an introduction to the concepts used in the thesis and to introduce the reader to the research area. Section 8 contains a summary of the research papers in Part II and III. The Appendix contains the search topics used in the tests in the thesis.

2 Linguistic features in information retrieval

Linguistic features in general and language specific features in particular are important in cross-language information retrieval. In monolingual information retrieval, Sheridan and Smeaton (1992) and Strzalkowski (1996) have incorporated into retrieval systems and at a relatively early stage among others linguistic techniques.

In this section linguistic features are discussed mainly to the extent relevant for the following empirical studies in the thesis. From this point of view the linguistic features considered important are:

- *Morphological variations - derivation - inflection*
- *Compounds and phrases*
- *Lexical ambiguity - homonymy and polysemy*

The empirical studies in this thesis deal with four different languages in cross-language information retrieval: Swedish, Finnish and German used as source languages in the requests and English used as target language, the document language, into which the translated queries are matched with the documents. Swedish and German as source languages are Germanic languages and have similar linguistic features. Finnish again is a Finno-Ugric language with a totally different and extremely rich morphology. A similar feature between the three source languages is however the formation of compounds. The target language English is morphologically not so complex and the formation of compounds is different. Most English compounds are in the form of multiword phrases not orthographically written as one word as in Finnish, Swedish and German.

Two forms of lexical ambiguity, important especially from the cross-language information retrieval point of view are homonymous and polysemous words. Homonyms are different lexemes spelled similarly, while a polysemous word is a word that has several sub-senses (Teleman, Hellberg and Andersson 1999). In the case of homonyms the sample languages differ very much. Swedish is rich, while for example the Finnish language is scarce in homonyms (Karlsson 1994).

2.1 Inflectional and derivational morphology

Morphology is the area of linguistics dealing with the internal structure of words. A **word** is in this case a **lexeme** that can have several word forms, e.g., the word *write* can take the forms *writes, wrote* and *written*, usually called inflected forms. The **base form** (in this case *write*) is the form from which the other forms of the lexeme can be derived using the morphological rules of a language. A **stem** is the form to which the inflectional suffixes are attached. (Lyons 1981)

A **morpheme** is the smallest unit of a language that has a meaning and cannot be broken down further into meaningful or recognisable parts. On a sub-word level, a word (lexeme) is constituted of either **bound** or **free morphemes**. A free morpheme can appear as an independent word in a sentence, e.g., the word *tree* is an independent word in the sentence, *I saw a tree*. A bound morpheme is always attached to another morpheme, as the morpheme *s* denoting plural in the word *trees*. Certain bound morphemes are called **affixes** and can be classified into **prefixes** and **suffixes.** A prefix is attached to the beginning of another morpheme, e.g., *re-* in *restore,* while a suffix is attached to the end of a morpheme, e.g., *modern - modernise*. (Akmajian, Demers, Farmer & Harnish 1995)

In many Germanic languages compound words are usually formed by joining two or more words to one orthographically. The **joining segment** is a morpheme, which in Swedish is named "**fogemorpheme**" and in German "**Fuge-element**" (Malmgren 1994; Fleischer and Barz 1992). They can take several forms and the conventions of using them are irregular. For example, the "*s*" joining *Handel* and *Vertrag* in *Handelsvertrag* (trade agreement) is a joining morpheme in the German word. Joining compound components may also cause the omission of the vowel, e.g., "*a*" in the Swedish word *skola* (school) in the compound *skolhus* (school building). Only the stem *skol-* is used.

Morphology can be broken down into the subclasses of **inflectional** and **derivational morphology**. Inflectional morphology describes the predictable regular changes a word may undertake as a result of syntax, e.g., plural forms, verb conjugation, adjective inflection for comparison, gender, case etc., e.g., the adjective comparison *large - larger - largest.* Derivational morphology describes how affixes combine with word stems to derive new words. Derivational suffixes may affect the part-of-speech and meaning of a word, e.g., *build - builder - building.* (Akmajian et al. 1995)

The inflectional morphology for Finnish is particularly rich, e.g., a word can take as many as 14 different terms in the category of case (Karlsson 1994). In this thesis, in Part II, special attention is paid to Swedish morphology and its impact on information retrieval, since Swedish is a "new" language in research on cross-language and monolingual information retrieval.

2.2 Compounds and phrases

Compounding is in most languages a common way to form new words from the already existing ones. By joining two or more words together like in *jet-plane, sailboat* or *pine tree* new compound words may be formed. There is no limit to the formation of new words, therefore compounding is a very productive process, and there is no dictionary that can contain all possible new compounds that may be created. Many new and occasionally formed compounds are therefore missing even in comprehensive word lists, e.g., "Svenska Akademiens ordlista", a word list of the Swedish language (SAOL 1998).

Compound analysis has in literature often been performed from the point of view of linguistic description. For English, compound words have been analysed in the context of word formation by Bauer (1983), for compound nominals by Levi (1978) and Warren (1978). German compounds have been analysed in the context of word formation by Fleischer and Barz (1992). Swedish compounds have been analysed by Noreen (1904-1907) for word formation and meaning relations. From the point of view of computational analysis Swedish compounds have been analysed by Blåberg (1988). Karlsson (1992) provides a description of the general morphological analyser for Swedish. Finnish compounds have been analysed from the point of view of computational analysis by Koskenniemi (1983).

The computational treatment of nominal compounds is troublesome, and as Sparck Jones (1983) argues, interpreting compounds requires inference in an unpredictable way, for example in the attempts to characterise nominal compounds in terms of general semantic relations (Warren 1978; Levi 1978). According to Blåberg (1988) compounds are likely to be treated in future computational applications according to the purposes of the particular application in question. Alternative interpretations may be accepted for example for information retrieval purposes, while for translation applications explicit representations may be required.

The **orthography of compounds** may be different depending on language but also within a language the conventions may differ. A compound may be written with a hyphen, as one word, or as separate words. English is a typical language where the orthography is inconsistent, but where multiword compounds generally are written as separate words (Akmajian et al. 1995). German, Dutch, the Scandinavian languages as well as Finnish have a much more consistent convention for writing compounds. Compound words are generally written as one word, even if they are formed of many components, e.g. *Kriegesdienstverweigerer* (Ge), *vapenvägrare* (Swe), *aseistakieltäytyjä* (Fi), all meaning *conscientious objector*. The joining morphemes used in German and Swedish were mentioned above and will be described in greater detail in the studies on Swedish morphology and on compound handling for cross-language information retrieval in Part II.

In this thesis **compounds** form an orthographically united class, written without intervening spaces, while a **phrase** denotes a compound expression written as separate words. This orthographic specification is important in the cross-lingual studies in this thesis. Phrases are treated word-by-word in the source languages and if a phrase expression is the outcome of the translation dictionary it is treated as a phrase. The studies in this thesis focus on compounds, therefore the reader is referred to the studies in Part II for a more detailed description on compounds and the importance of correct handling of compounds and phrases in cross-language information retrieval.

What makes compounds interesting from information retrieval point of view is the possibility to create new words by compounding. These words are less likely to appear in translation dictionaries, and therefore dictionary-based cross-language information retrieval may be complicated. Since for many compounds only their constituents can be found in dictionaries, it is important to be able to **split compounds into constituents**. A particular case of compounding is that compounds can consist of other compounds, e.g., *Methangaslagerstätte* (deposit for methane gas) consists of the two compounds *Methangas* (*Methan* and *Gas*) and *Lagerstätte* (*Lager* and *Stätte*).

Compounds have in the literature been described as syntactic expressions, e.g., they are derived by grammatical rules (Blåberg 1988). If the meaning of a compound can be deduced compositionally from the meaning of its parts it is a **compositional compound** and the compound expression is a syntactic expression. If the meaning of the compound can be interpreted through the components, compound splitting is in general useful in dictionary-based cross-language information retrieval. E.g., *Weltwetter* can be decomposed into *Welt* (world) and *Wetter* (weather) and the components may be translated even if the whole compound is not found in the translation dictionary. Compounds also share properties with typical lexemes (Blåberg 1988). For **non-compositional** compounds like *Erdbeere* (strawberry) where the meaning cannot be interpreted through the components *Erd* (ground, soil) and *Beere* (berry), decomposing the compounds may add noise to the translated query.

However, in information retrieval the case where *berry* is a **hyperonym**, a "headword" for all kinds of berries, e.g., cloudberries, blueberries, is very common. Splitting the compound makes this last component (the head) searchable even if the whole compound would not have been translated. This is a very interesting property of compounds and may also be used to expand queries in information retrieval (Pirkola 1999).

For the source languages used in this thesis, Swedish, Finnish and German where multiword expressions are compounds rather than phrases, translation of phrases for cross-language information retrieval is not a major problem (Pirkola 1999). However, the identification of phrases has been regarded as an improvement in retrieval performance for Spanish to English (Ballesteros & Croft 1997).

2.3 Lexical meaning and ambiguity

Three forms of **lexical meaning** (interpreted as the meaning of lexemes), are relevant for information retrieval: 1) ***homonymous*** 2) **polysemous** and 3) **synonymous** words.

Two words (lexemes) are homonymous if they have at least two distinct meanings and the senses of the words are unconnected, e.g. *bank* (financial institution) and *bank* (river bank). A lexeme with more than one sense is polysemous, e.g. *star* in the sky and *star* (a famous person). The senses of polysemous words are related to each other. Synonymy again is defined as identity of meaning between two lexemes, e.g. *aircraft*, *aeroplane.* (Lyons 1981)

From the information retrieval point of view, lexical ambiguity covers homonymy and polysemy (Pirkola 1999). Due to ambiguity in the search keys, matching may not be successful for retrieving relevant documents. Some languages, like Swedish, are very rich in homonymous words, around 65% in running text, while for Finnish only 15% are homonyms (Karlsson 1994).

Homonymy and polysemy are recognised by lexicographers in translation dictionaries, but the distinction between them is hard to apply in a consistent way. The relatedness of meaning as a condition for polysemy can be derived from historical or etymological reasons, but the line is hard to draw (Malmgren 1994; Lyons 1981).

The phenomenon of **translation ambiguity** is common in cross-language information retrieval and refers to the increase of irrelevant search key senses due to lexical ambiguity in the source and target languages (Pirkola, Hedlund, Keskustalo & Järvelin 2001). A search key may have one or several senses in the source language, which in the translation dictionary are expressed by several translation alternatives. In the translation process to the target language extraneous senses may be added due to the fact that each translation alternative may have several senses. Thus lexical ambiguity in cross-language queries appears both in the source and target language.

Synonyms have an identity of meaning to a certain degree, from absolute synonyms if they are synonymous in all the contexts they appear in, to synonyms in a certain range of context (Lyons 1981). Absolute synonymy is very rare and normally we talk about synonyms in quite a broad sense.

The translation alternatives listed in a dictionary are naturally also mostly synonyms and therefore expand the query if they are accepted as translations into the final query.

3 Natural language tools for information retrieval

Natural language processing means that natural language texts are analysed automatically for the purpose of information retrieval, automatic translation, text generation etc. The aim in natural language processing research is to create robust systems that can handle large numbers of text documents in a reasonable time (Haas 1996). Natural language processing covers both statistical and linguistic methods. Different levels of **linguistic analysis** can be identified:
1) **morphological,** where the structure of words is analysed, 2) **syntactic**, covering the structure of sentences, 3) **semantic,** involving the meaning of words and sentences and 4) **discourse** analysis, where texts are analysed in their textual context. Stemmers and normalisers as examples of morphological tools will be discussed in the subsequent sections. The use of stop word lists, that is, the removal of frequent non-significant words in requests and documents is also mentioned as important in information retrieval. The last section will briefly discuss other forms of linguistic analysis from information retrieval point of view.

3.1 Stemmers

The most common morphological tools for information retrieval applications are **stemmers** for producing word stems and morphological **normalisers** (lemmatisers) for normalisation and compound splitting. Stemming, or truncation of suffixes, was also one of the first approaches to connect morphology and information retrieval (Salton & McGill 1983).

Stemming is a computational process removing inflectional and derivational affixes and returning a word stem, not necessarily a real word. The main difference to morphological **normalising** is that normalising turns the word to its lexical full base form. The negative effect of stemming and normalisation as processes in information retrieval is that they may produce noise as unrelated word forms are sometimes conflated to a single form.

Stemming is traditionally considered to improve recall in information retrieval systems since more potentially relevant documents can be retrieved. The effect of stemming on precision is more controversial. For the English language, Harman (1991) tested three linguistic stemmers but found very little improvements to the retrieval effectiveness. On the other hand, Hull (1996) and Krovetz (1993) found that stemming improved precision in English. The most well known stemming algorithms, the Lovins and Porter stemmers (Lovins 1968; Porter 1980), are suitable for morphologically simple languages like English where few stem changes occur (Tzoukerman, Klavans & Jacquemin 1997). The results on English may not be directly applicable to other inflectionally more complicated languages. English is a typologically special language in the sense that word order is more important than inflection (Karlgren 2000).

Savoy (1999; 2002) has developed a "quick and dirty" stemmer for French, tested on medium-sized French collections. Based on the same concept, stemming algorithms for Italian, Spanish and German were implemented and used in cross-language retrieval with good results for Spanish and Italian but less good for German. For morphologically more complex languages, especially languages where compounds need to be decomposed (German, Dutch, Scandinavian languages and Finnish), a linguistically more complex stemmer is needed. Kraaij & Pohlman (1996) compared the Porter-style stemmer to linguistic stemmers (both derivational and inflectional) for Dutch and the best results were achieved by an inflectional stemmer combined with compound splitting. Applying both inflectional and derivational stemmers generally reduces precision too much. Carlberger, Dalianis, Hassler & Knutsson (2001) have developed a stemmer for Swedish and tested it on Swedish documents. Information retrieval results with stemming were better than retrieval without stemming. Alkula (2001) compared stemming and word normalisation with regard to retrieval performance for Finnish. Gey, Jiang, Petras & Chen (2001) report that the experiences from the Spanish tracks in TREC are that some form of stemming will always improve performance, and from CLEF experiments that language specific stemmers result in an improvement in automatic multilingual retrieval.

Ripplinger (2001) argues that in German, standard stemmers (Porter) have serious deficiencies since they perform stemming by simply chopping off suffixes. This procedure results in word stems, not lexical base forms. In recent German studies Braschler & Ripplinger (2003) report experimental results from tests using stemming and splitting of compound words for German monolingual retrieval. Their main findings are that stemming in most cases is beneficial for German text retrieval and that decompounding contributes more to the performance than stemming. The results on stemming for German confirm the results by Tomlinson (2002) among others. Advanced stemmers have been developed and combined with a lexicon to verify the identified form (Krovetz 1993). This is a feature that has very much in common with morphological normalisers.

Automatic stemmers, applicable on more than one language, are a challenge in an environment with multiple languages such as cross-language information retrieval. Xu and Croft (1998) tested an automatic trigram stemmer on Spanish and English with comparable results to the performance of the Porter and KStem algorithms. Language-independent techniques have also been tested by Oard, Levow & Cabezas (2001), simplifying morphological analysis by constructing simple statistical stemmers based on word statistics of a text collection. However, although the statistical stemmer was performing well for the French language, the compound-rich language German needs compound splitting in order to obtain good results.

3.2 Normalisers

Much of what is said about stemmers and their implications on information retrieval can be transferred to morphological normalisers. The main difference is the capability of returning lexical base forms. Still the quality of the analysis of all morphological analysers, both lexicon-based stemmers and normalisers, depend on the size of the lexicon. In morphologically complex languages, morphological normalisers are needed especially for cross-language information retrieval. In dictionary-based cross-language information retrieval a lexical base form of a word is needed in order to match the entries in a translation dictionary.

For compound rich languages, compound decomposition is an essential feature, because of the problem with embedded search keys. If compounds are not decomposed, the non-first components are not retrievable. The last component is often a hyperonym of the full compound (Pirkola 1999). Different types of berries *blueberry, strawberry, cloudberry* all have the common hyperonym *berry* as the last component. If the compounds are split into constituents, one search key *berry* covers all types of berries. Sophisticated morphological analysers can also inform of inflectional and part-of-speech categories for the analysed word.

In the experiments included in this thesis, morphological analysers for Finnish, Swedish, German and English were used. The TWOL analysers performed normalisation and compound splitting and are based on the two-level morphology by Koskenniemi (1983). English was used as a document language and in the indexing phase of the document database the same morphological analyser was applied.

3.3 Handling of joining morphemes

A specific feature in many Germanic languages (German, Dutch and the Scandinavian languages) is the use of a joining morpheme in compounds. For example, the German compound noun *Handelsvertrag* (trade agreement) has two constituents *Handel* (business, trade) and *Vertrag* (contract, agreement) which are joined by the joining element "*s*". Morphological analysis tools do not necessarily remove the fogemorpheme when splitting compounds into constituents. That is, they do return the first constituent as *Handels* and not as the lexical base form *Handel.* In the automated process described and empirically tested in this thesis, an algorithm was constructed to handle fogemorphemes when splitting compound words in Swedish and German.

3.4 Stop words

The process of removing frequent non-significant words (stop words) in a document or a request is normally done using so-called stop word lists. Stop word lists have been used in monolingual information retrieval systems for the removal of high frequency words like prepositions, articles, pronouns, conjunctions, common verbs etc. The same function of removing non-significant words is needed in cross-language retrieval as well. Some general guidelines for stop word lists are found in Fox (1990). Savoy (1999; 2002) has developed a stop word list for French following the general principles of Fox. The stop word list has been extended to other European languages in cross-language experiments.

For the experiments in the studies on cross-language research in this thesis, stop word lists for English, Finnish, Swedish and German were used. The English stop word list was the one provided in the information retrieval software (InQuery). For a description of the software see Section 6. To establish stop word lists for Finnish, Swedish and German, the English list was translated to the respective languages using bilingual dictionaries. The translated lists with in some cases several translation alternatives for the original word were then modified to suit the needs of that particular language. The topics provided by the Cross-Language Evaluation Forum (CLEF) that were used in the tests in this thesis contain repeated expressions like "relevant documents contain". The words in such expressions were added to the stop word lists.

3.5 Other forms of natural language analysis in information retrieval

In the above mentioned morphological processes, base forms are produced of words handled one-by-one without their context. However, as Strzalkowski (1999) argues, bag-of-word representations support content-based information retrieval insufficiently.

A **syntax level analysis** considers sentences and the relationships between words in a sentence. More commonly, grammar defines valid relationships between words. A syntax level analysis for information retrieval purposes would be able to determine head-modifier relationships for example in noun phrases. Phrase identification in some form is according to Strzalkowski (1999) probably one of the most popular linguistic analysis forms in information retrieval applications. As an example of this type of analysis, Mitra, Buckley, Singhal & Cardie (1997) report interesting work on syntactic and statistical phrases. A statistical phrase is defined as any pair of words that occur contiguously frequently in a text corpus and a syntactic phrase is formed by any sequence of words that satisfy certain syntactic structures. The hypothesis is that syntactic phrases are able to express the semantic meaning in a better way. However, the results of the tests by Mitra et al. show that syntactic and statistical noun phrases yield comparable performance. The impact of phrases on a basic good ranking system did not effect the order of highly ranked documents but it is found useful in the set of low ranked documents.

In the information retrieval context, the need for one single correct analysis for every language construct is not present. It is sufficient to encode the possible interpretations of a construct so that it is available for matching in the retrieval phase (Sheridan & Smeaton 1992). Syntactic level processing is domain-independent, but by using only such level processing in information retrieval the results, according to Sheridan & Smeaton, have not been very promising.

Semantic processing attempts to identify the meaning of words in a sentence. It is a very complicated task and heavily domain dependent and requires world knowledge. Incorporating semantic-level processing into retrieval has led to conceptual information retrieval, which is effective but domain specific (Sheridan & Smeaton 1992).

The depth of natural language analysis can be relatively shallow and still be able to improve the representations of text in indexing and requests compared to string-based methods in statistical retrieval (Stzralkowski, Lin, Wang & Carballo 1999). In the tests made by Strzalkowski et al. the success of natural language analysis in information retrieval was found to be related to query length. Long and descriptive queries seem to respond well to natural language processing while shorter queries show hardly any improvement.

4 Cross-language information retrieval - problems and approaches

In cross-language information retrieval research the main test and problem situation is to present a query in one language against a document collection in another language and by filtering, selecting and ranking documents produce a result relevant to the request (Grefenstette 1998). A query is the request expressed as search keys in a form that the retrieval system is able to process. Search keys are the expressions selected to represent the request for the information retrieval system. Despite the language aspect, cross-language methods and monolingual methods have much in common and monolingual research results can to some extent be adapted to and help us understand cross-language retrieval research problems. In traditional information retrieval, the focus has been a character string matching approach rather than a natural language processing approach, which pays more attention to the nature of texts. As a consequence, the role of language resources in standard information retrieval system has remained marginal (Gonzalo 2001). However, as Gonzalo states, in cross-language information retrieval the language aspect where queries are presented in different languages is "changing the landscape". Cross-language information retrieval must combine linguistic techniques with robust monolingual information retrieval (Gey et al. 2001).

Historically, the method for cross-language retrieval with a controlled vocabulary is the oldest one (Salton 1970), and also used in operational systems like library catalogues. The cross-language approach using controlled vocabulary involves the translation (indexing) of both documents and the queries to a common language, that of the controlled vocabulary. The translation of terms is done by using a bi- or multilingual thesaurus, which relates the terms from each language to each other, or to a common language-independent set of identifiers (Oard 1997). The project EuroWordNet, developing a database of WordNets for a number of European languages, has constructed a structure similar to the controlled vocabulary thesaurus used by Salton (Gollins & Sanderson 2001; Vossen 1997).

After the experiments with controlled vocabulary the research community focused on the free text approach. This was naturally due to the growing number of texts available in electronic form. Also public forums for the evaluation of research results, e.g., the Text Retrieval Conference (TREC) started with a cross-lingual track in 1997 (TREC 6). In the year 2000, as a continuation and expansion of the cross-lingual track as a part of TREC, a workshop for cross-language evaluation (CLEF) was held in Lisbon, Portugal (Peters 2001). CLEF concentrates on European languages, hoping to increase the knowledge of non-English resources. A similar continuation to TREC for Asian languages is the NTCIR workshop (Kando 2001).

The first question is to consider either translating the query or translating the documents. Since document translation is expensive it seems obvious that in most cases it is more effective to initially translate only the query. Interesting documents retrieved can be judged, titles and abstracts can be roughly translated to the user, if necessary, before the actual decision on which documents should be completely translated.

The two main approaches to free text cross-language information retrieval are methods based on stored external knowledge (knowledge-based methods) or methods based on the analysis of text corpora (corpus-based methods). This distinction is becoming less useful for classification of systems since merging of available resources is becoming more common (Gonzalo 2001).

The **corpus-based approaches** start from text analysis. Document text collections in different languages form the text corpora needed for this approach. The aim is to extract the information needed for the translation from the existing texts. The text collections can include exactly the same texts in several languages (parallel corpora) or the texts can include documents belonging to the same subject category (comparable corpora) (Oard 1997). Relevant documents in the source language are retrieved and words are extracted from parallel or related documents in the target language. Approaches like cross- language latent semantic indexing apply a mathematical matrix suppressing technique (singular value decomposition) to compose vectors expressing the document content. A mapping function creating short dense vectors is the output, which suppresses the term usage variation (Landauer & Littman 1990; Littman, Dumais & Landauer 1998).

The **knowledge-based approach** is based on knowledge structures. They can be in the form of multi- or bilingual dictionaries or thesauri applied to free text retrieval, or in the form of sophisticated ontologies, e.g. the EuroWordNet.

Of the knowledge-based approaches, the most thoroughly explored branch is the **dictionary-based method**, which relies on standard bi- or multilingual dictionaries that are transformed into a machine-readable form. This approach offers a relatively cheap and easily applicable solution for large-scale document collections. Dictionaries are used to translate each word of the source language query to the desired target language. In the translation process words can be translated by not one unique term but a set of terms appearing as equivalent translations in the dictionary. The ambiguities arising in the translation phase are understood and described using the linguistic concepts of polysemy and homonymy. There is a need to disambiguate homonymous and polysemous words. The need is greater in cross-language retrieval compared to monolingual information retrieval. Therefore the linguistic features in a language become important and cross-language systems very often would benefit from tools for natural language processing.

Morphological analysis tools, able to normalise inflected word forms to base forms and to decompose compounds into constituents, may be used in cross-language information retrieval systems in the index building phase as well as in translating queries. Normalisation is utterly important in dictionary-based cross-language retrieval to be able to match query words to dictionary entries and in matching translation output to the database index.

Machine translation is another linguistic and knowledge-based approach available for query translation. However, machine translation systems are able to produce high quality translations only in limited domains (Oard & Dorr 1996). They need information about context and are based on syntactic analysis. Syntactic analysis is not possible for the translation of bag-of-word queries, lacking grammatical structure. However, machine translation has been used as a method in some research reports on cross-language retrieval. The applications include translations of documents (Davis & Ogden 1997), a front-end tool for cross-language information retrieval applications (Yamabana, Muraki, Doi & Kamei 1998), and an approach to use the technology of an existing machine translation system SYSTRAN for query translation (Gachot, Lange & Yang 1998; McNamee, Mayfield & Piatko 2001).

In the following sections research using dictionary-based methods will be looked at more closely, since these methods are used in the following empirical studies in this thesis.

4.1. Dictionary-based methods

The main problem areas occurring in dictionary-based cross-language information retrieval are defined in Pirkola et al. (2001):

- Untranslatable search keys due to limitations in dictionaries.
- Processing of derived or inflected word forms.
- Phrase and compound translation.
- Lexical ambiguity in source and target languages.

The problem areas will be discussed below from a general perspective as well as their implications to the research in this thesis.

Untranslatable search keys

Not every word form used in a text or query is always found in a dictionary. This may be due to the domain of the query. A specific domain can develop a specific terminology not included in a general dictionary. The translation problem may also be due to **compound words**. Compounding is a way of creating new words and therefore their proper translation is not necessarily included in dictionaries. The problem with identifying **proper names** in a query (Pfeifer, Poersch & Fuhr 1996) exists for names of persons, places, and organisations. In languages with rich inflection proper names may also appear in inflected forms in text and therefore the identification and matching of the same name in a text in a different language is difficult. Names of persons, places, organisations would need to be normalised and translated but proper names are not generally found in the lexicon of a morphological analyser and translations of names seldom appear in a dictionary. Also when transliterating names from for example Russian or Chinese different spelling conventions are used (Grefenstette 1998). However, important cities and country names are normally included in dictionaries. In cross-language information retrieval systems a word not recognised by the dictionary is typically added to the target query in the form it appears in the source language query.

In the research papers included in this thesis (Article 2 in Part II and Article 2 and 3 in Part III) a method based on approximate string matching is used to solve the problem with proper names. The algorithm was developed and described by Pirkola, Keskustalo, Leppänen, Känsälä & Järvelin (2002).

Processing of derived or inflected word forms

Inflected word forms are usually not included in dictionaries. Normalisation or stemming of word forms in a query is therefore an important step in a cross-language information retrieval system. Fluhr et al. (1998) use linguistic analysis as a base for normalisation and identification of compounds. Part-of speech identification is done using corpus-based syntactic knowledge. Fluhr et al stress the importance of high quality in linguistic analysis and particularly in the treatment of compound words and phrases. Having source words in base forms makes them easily translatable. However, normalisation is not an unambiguous process - source tokens may have several possible base forms due to homography and polysemy. This introduces the need for word sense disambiguation (Krovetz 1993).

In the research paper in this thesis, normalisation of source language words was included in the features of the UTACLIR query translation system. For the normalisation, externally developed morphological analysis tools were used. However, for compound splitting and normalisation of compound components in Swedish and German an algorithm handling the joining morphemes in compounds was developed.

Phrase and compound translation

Identification and use of phrases in information retrieval is traditionally considered important (Croft, Turtle & Lewis 1991; Buckley, Singhal, Mitra & Salton 1996). Statistical methods and syntactic analysis to identify phrases has been discussed above in Section 3.5.

Proper translation of phrases is also important in dictionary-based cross-language information retrieval. Earlier studies by Hull & Grefenstette (1996) for French - English cross-language retrieval found that compared to monolingual information retrieval individual components of phrases have very different meaning in translation. In their study manually built dictionaries containing phrases increased the performance compared to word-based dictionaries.

The problem with phrase translation is not crucial in languages where multi-word expressions are compounds rather than phrases (Pirkola 1999), instead proper translation of compounds is important.

In this thesis 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 a language where multi-word expressions are phrases.

In this thesis, Part I article 2 and Part II articles 1,2 and 3 deal with the problem of compound handling. As a result of the research it was found that for compound-rich languages compound splitting in the source language generally seems to improve performance. In this thesis where the target language is English, a non-compound language, settings allowing a phrase structure in the translated target language query were thought to be the best solution.

In the target language a **phrase-based structuring of compounds***,* imitating a phrase in the target language was performed since word-by word translation of compound components does not automatically support a phrase structure in the target language. The translation equivalents that correspond to the first component of a compound were joined by a proximity operator to the translation equivalents of the second, third etc. component. All the combinations were generated. [See Section 6 for a description of InQuery's synonym and proximity operators] For example, the Swedish compound "*mötesplats*" (place of a meeting) is decomposed into "*möte*" and "*plats*". For the first component, a translation dictionary from Swedish to English could give the translation equivalents (meeting, date, appointment) and for the second component (place, room). A phrase-based structuring of the source language compound would give the following combinations joined by a proximity operator, here (#uw):

> #uw(meeting place), #uw(date place), #uw(appointment place), #uw(meeting room) #uw(date room), #uw(appointment room)

The proximity statements are combined by a synonym operator, here (#syn):

> #syn(#uw(meeting place), #uw(date place), #uw(appointment place), #uw(meeting room) #uw(date room), #uw(appointment room)).

However, the findings indicate that when the proximity operators in the translated target language query was substituted by synonym operators, the results were beneficial for the latter query setting. On individual query level the results varied. An explanation to the results can be found in earlier studies on English monolingual retrieval, for example, Mitra et al. (1997). Their findings were that, even though phrases are considered to improve precision, adding a phrase structure to a query that already achieves good performance using single terms, can over-emphasize a particular aspect in the query. In a later study by Pirkola, Puolamäki & Järvelin (2003) the effect of the use of the *uw* proximity operator was tested. Their findings were that the phrase-based structuring of compounds was not helpful for synonym structured queries.

Lexical ambiguity in source and target languages.

In cross-language information retrieval we are not restricted to use only one translation alternative like in machine translation. Several translation alternatives can be added to the query. The problem that occurs is however that of choosing among and weighting these alternatives. If we have a query containing four words that should have equal weight in the query and translate the first one of them with one unique word, the second with three alternative equivalent translation alternatives, the third one with as many as six alternatives and the last one with two alternatives, we accidentally add more weight to the words with several translation alternatives (Grefenstette 1998).

Structuring the queries *in a syn-based structuring* where alternative translations are connected by a synonym operator is a method developed by Pirkola (1998) and presented and tested by Pirkola et al. (2003) with positive results. The Pirkola-method has been applied in the research papers of this thesis, using the *#syn* operator of the InQuery retrieval system. The translation equivalents of a source language key are grouped by the *#syn* operator. For example, the Finnish word "*tauti*" (illness) might get the following translations in a dictionary (illness, ailment, disease). The syn-based structuring groups the translation alternatives using the syn-operator in the following way:

#syn(illness ailment disease).

Hull (1998) tested a weighted Boolean model for CLIR. This method ensures that the number of translation alternatives does not influence the ranked query result more than the original query term would influence the ranking of source language documents. Sperer & Oard (2000) tested structured queries as a solution to ambiguity problems and the tests found the Pirkola-method very effective. The effect of query structuring in automated cross-language information retrieval is also tested in this thesis.

Translation polysemy involves the fact that most languages contain polysemous expressions, that is, a word (lexeme) can have several meanings depending on its context. Adding translation alternatives that are not equivalent or synonyms to the meaning of the words in a source query add noise to the translated query. Ballesteros & Croft (1998a) reported that the transfer of inappropriate senses to the query causes a loss of effectiveness and is a bigger problem for longer queries. Different kinds of filtering and disambiguation methods have been tested in cross-language information retrieval research. One way, used at least by Hull (1998) and Fluhr et al. (1998), is to let the text collection perform the filtering. Davis (1998) executes a disambiguation process where the translation equivalents accepted into the query are selected from a parallel corpus. Thus, the research by Davis is also an example of the combination of dictionary and corpus-based methods.

4.2 Research results in cross-language information retrieval

Already the results achieved by Salton (1970), were favourable. The system, which used a manually constructed bilingual thesaurus performed almost as well as monolingual retrieval. However, the test collection was small and today it seems unrealistic to manually index a large document collection. Also automatic thesaurus construction needs more development (Oard & Dorr 1996).

For cross-language systems a good benchmark system that gives the upper bounds of performance is a monolingual test under the same experimental condition (Oard & Dorr 1996). See Figure 1 for a test setting for the UTACLIR experiments in this thesis.

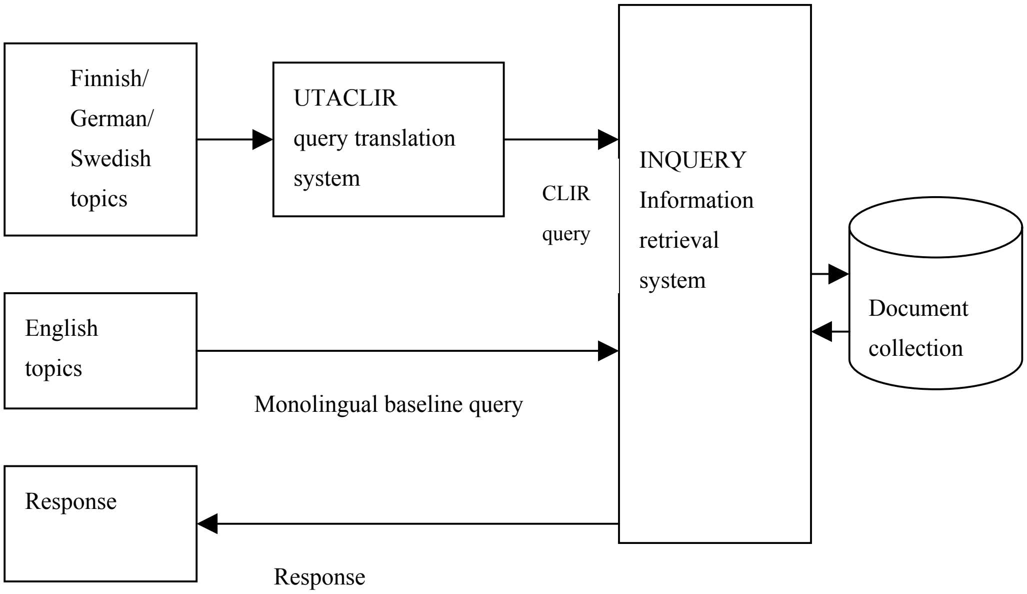


Figure 1 Test setting for the UTACLIR experiments

The cross-language results tend to be in a range of 50% to 75% of the equivalent monolingual runs (Harman et al. 2001). Research methods using bilingual dictionaries as the only method tend to perform in a range of 40-60% below the corresponding monolingual retrieval result (Ballesteros & Croft 1998a; Hull & Grefenstette 1996). According to Ballesteros and Croft the many extraneous words in a translated query account for a clear loss in performance, and this increases with longer queries. A second factor that causes loss in performance is the translation of phrases as individual words.

By using combined techniques Ballesteros and Croft (1998a) found that cross-language information retrieval systems tend to perform better. For example, adding corpus-based feedback prior to dictionary translation improves precision especially of short, less specified queries while corpus-based feedback and query modification after the translation process improves recall. Pirkola (1998) shows that a combination of query structuring and the simultaneous use of general and special domain dictionaries yields an improvement in performance of cross-language information retrieval queries to nearly that of monolingual queries. Disambiguation using a parallel corpus can reduce the performance drop with respect to monolingual queries. Davis (1998) shows this relation in his test and finds that a complete translation using dictionaries causes a 58% drop in average precision from monolingual queries, but the disambiguated queries only show a 33% drop. A part-of-speech tagging could still improve the performance. Hull argues that the ambiguity associated with language translation using dictionaries could be automatically resolved using a weighted Boolean model (Hull 1998).

Ballesteros & Croft (1998b) use a combined method of dictionary and co-occurrence statistics for disambiguation. They focus on the correct translation of phrases. The hypothesis is that correct translation will co-occur in sentences and that incorrect translations will tend not to. Unlinked corpora can in this case perform as a disambiguation method as well as parallel or comparable corpora. They also find that the use of a synonym operator for translations having more than one target language equivalent is more effective for disambiguation than part-of-speech tagging.

The use of the latent semantic indexing method does not involve dictionaries so the typical problems of dictionary translations do not exist. However, even though the tests performed by Littman et al. (1998) showed a good performance the method is at an experimental stage.

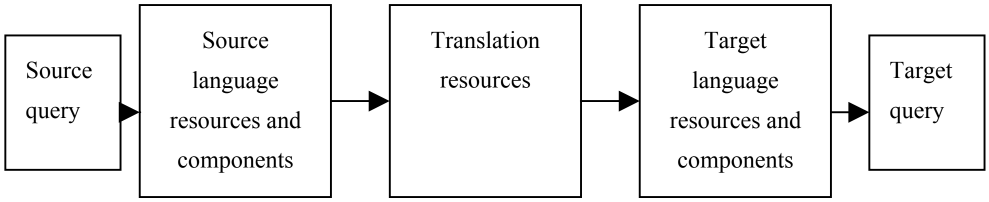
The languages represented in cross-language information retrieval research are mostly large languages like English, Spanish, French and Chinese. However, also smaller European languages like Finnish and Swedish have been used especially in the CLEF experiments. Some of the main characteristics of the CLEF experiments are that a majority of participants, 14 of 20, used a dictionary-based method either as the core method or in a combination with corpus-based methods or / and machine translation. The top-performing runs in the multilingual task used a combination of translations from multiple sources (Braschler 2001). The different tasks, multilingual, bilingual and monolingual (non-English), were all judged separately. The five best results in each task were reported in the overview of the results in the proceedings (Braschler 2001; Peters 2001). The UTACLIR system by the University of Tampere performed well in the bilingual task both in the CLEF 2000 and 2001 evaluation campaign coming in at a fourth respectively fifth place. The result however does not take consideration to which language pairs were used or other experimental settings, only best performing runs measured by average precision. Compound splitting also seemed to increase the performance at least among the five best performing runs in the bilingual task for compound-rich source languages (Hedlund, Keskustalo, Pirkola, Sepponen & Järvelin 2001; Hiemstra, Kraaij, Pohlmann & Westerveld. 2001) as for the monolingual German task (Moulinier, McCulloh & Lund 2001).

5 The UTACLIR query translation system

5.1 The general UTACLIR framework for query construction

The UTACLIR query construction framework is an automated dictionary-based process where we consider specific linguistic features of source and target language words. The features chosen, morphology and compounds were identified as important in cross-language information retrieval. The process has been developed, improved and evaluated for three CLEF campaigns in 2000 and 2001 and 2002 for bilingual dictionary-based cross-language retrieval. The latest version of the software was presented at the Workshop on cross-language information retrieval of the ACM SIGIR Conference in Tampere in 2002 (Hedlund, Keskustalo, Airio & Pirkola 2002).

A rich inflectional morphology requires word form normalisation in order to match source keys with dictionary entries also in the case of inflected source keys. Thus morphological analysers are relevant in our framework. Since a lexicon of a morphological normaliser and a translation dictionary never holds all words present in a language, also word forms not recognised by the linguistic tools must be handled by the system. For each source language we utilise word form normalisation, the removal of stop words and handling of untranslated words. For compound source languages also compound splitting and normalisation of compound components were supported. The target language query formation supports normalisation, synonym structuring of translation equivalents and phrase-based structuring of target phrases (see Figure 2). The process is described in detail in the papers in Part III of this thesis. 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. (2002).


Source query
Source language resources and components
Translation resources
Target language resources and components
Target query

External resources:	**External resources:**	**External resources:**
normalisation tools	bilingual dictionaries	normalisation tools
stop word lists		stop word lists
System components:		**System components:**
compound handling		structuring of queries
fogemorphem algorithm		phrase structure for compounds
n-gram algorithm		

Figure 2. UTACLIR external resources and main system components

5.2 General description of the UTACLIR system

Process level

On a process level the UTACLIR system input is a structured or unstructured source language query, together with codes expressing the source and target languages. On the basis of this information query translation proceeds by utilising the general UTACLIR framework for translating the individual source keys using available linguistic resources. The processing of source query keys is based on seven distinct key types. The key type of the input word depends on the lexicon of the morphological analyser, the stop word lists used and the translation dictionary (Hedlund et al. 2003). The examples below are from the empirical studies in this thesis where compound languages are used as source languages.

In case the source key is recognised by morphological software:

1. Keys producing only such basic forms which all are stopwords.
   For example, the German word *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* (currency)
3. Keys producing no translatable basic forms.
   For example, proper names like *Pierre, Beregovoy, GATT*,
4. Untranslatable but splittable compound words.

For example, the German compound words *Windenergie,* decomposed to *Wind* and *Energie* (English translation: wind energy)

In case the source key is not recognised by morphological software:

1) The key is a stop word.
   There are very few examples for this key type, in the German topics the word *ex* is a key that could be a stop word.
2) The key is translatable.
   The German word *jordanisch* (Jordanian)
3) The key is untranslatable by the dictionary used.
   For example, the German words *Tschetschenien, Bosnien*

In the present implementation, all source keys recognised as stop words are removed first. All translatable source keys (also decomposed compound words) are then translated by using a translation dictionary.

System architecture

The UTACLIR system components and a system overview are presented in Figure 3. The input is a source language query; processed word by word and the output is a structured target language query. The activities; normalisation, stop word removal, translation, compound handling, translation of compound components, n-gram handling, target word normalisation and finally the structuring of the query are shown in the activity boxes. The external resources and the system components are arrows appearing from the bottom of the figure. Only in the case when the source key is not recognised by the morphological analysis tool and not able to translate by the dictionary used the n-gram handling is performed. Compound handling is performed only in the case where direct translation is not possible. The internal flow of the system is shown using arrows connecting the activities.

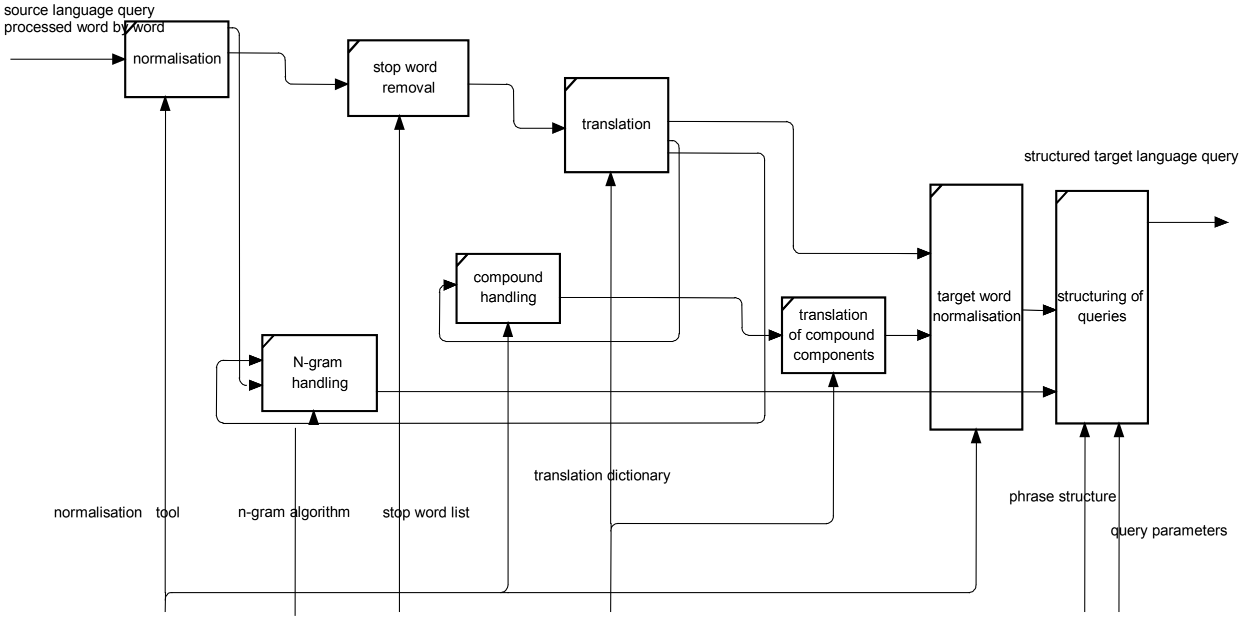


Figure 3. UTACLIR system architecture

System implementation

Internally 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 of nodes contain processed source language strings, for example words generated by morphological programs (for example, basic forms or parts of a split compound). Also word analysis information may be saved into 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 in this way, and given as the final output of the translation process. Additionally, analysis information (from the second level tree nodes) can also be the output. (Hedlund et al. 2002)

In the target query formation phase, key normalisation, synonym structuring of translation equivalents, and phrase-based structuring of target phrases are supported. For the remaining untranslatable keys, novel n-gram based techniques can be utilised to find the best matches from among the target database index words (Pirkola et al. 2002).

The system operates on Solaris 7 work station. It is programmed in C and consists of a library archive containing general and resource specific functions.

General functions are called to implement the basic translation service. Resource specific functions (e.g. interfaces to local dictionaries) and general functions (e.g. new ways to structure the target query) can be added, by writing functions satisfying the function prototype definitions given by the general system framework. Morphological analysis and word translation is performed by using the library archives of the respective external resource (morphological normalisers, dictionaries). Presently, UTACLIR utilises external language resources in a uniform manner by always calling for a simple data structure consisting of a linked list of word nodes. (Hedlund et al. 2002)

6 The InQuery Information Retrieval System

The basic retrieval models for text retrieval are based on three different approaches. The *Boolean retrieval model* is based on Boolean logic, the *vector retrieval model* is based on document and query representations expressed as weighted vectors in a vector space, and the *probabilistic model* is based on Bayesian inference networks. The Boolean systems are often called exact match systems, and they are based on binary matching, where documents either fully match a query and are retrieved, or do not match it at all and are not retrieved. Systems based on vector and probabilistic models are called best-match systems. They do not use a strict binary matching. Instead documents are ranked according to a ranking algorithm, computing the probability of relevance of a document to a query or some other matching score.

For the empirical tests in Part II and III, InQuery, a retrieval software developed at the University of Massachusetts, is used. InQuery is a probabilistic information retrieval system that uses a Bayesian network model to describe how text and queries should be used to identify relevant documents (Broglio, Callan, Croft 1994). In the probabilistic as well as in the vector retrieval model, documents are scored with respect to the query. As a basis of the ranking the $tf \times idf$ indexing weight (tf = frequency of a key in a document and idf = the inverse frequency of a key in the whole document collection) is used in both models. In InQuery, as in other probabilistic systems, the documents are ranked in a descending order of probability of relevance to the entered query.

For the InQuery (3.1) system a modified $tf \times idf$ weight is used at indexing time, and all keys are attached with a belief value computed in the following way (Allan et al 1997).

$$0.4 + 0.6 \times \left[\frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 \times \left[\frac{dl_j}{adl} \right]} \right] \times \left[\frac{\log\left[\frac{N + 0.5}{df_i} \right]}{\log(N + 1.0)} \right]$$

tf_{ij} = frequency of search key i in document j
df_i = the number of documents that contain the search key i
dl_j = the length of the document j
adl = average document length in the collection
N = number of documents in the collection

The document length variable dl_j and the variable expressing average document length *adl*, are used in the formula for document length normalisation. Without normalisation, the weights would favour long documents. If a search key is not occurring in a document, the default value 0.4 is assigned to the key.

The weight assigned to a document also depends on the types of query operators used. In the experiments for this thesis the following operators were used: the *sum* operator, the *syn* operator, the *od* operator and the *uw* operator. The *sum* operator is the default operator and used automatically in InQuery when no operator is specified. For the *sum* operator, the weight is computed as the average of the weights of the operand facets (the query search keys). The *syn* operator was used in the experiments in this study for the expression of a synonym structure of translation equivalents in the target query (Pirkola 1998). The *syn* operator was applied to the search words of a single facet, thus treating the operand search words as instances of the same search word. For the *syn* operator, the following key weighting formula is used (Kekäläinen, Järvelin 1998)

$$0.4 + 0.6 \times \left[\frac{\sum_{i \in S} tf_{ij}}{\sum_{i \in S} tf_{ij} + 0.5 + 1.5 \times \left[\frac{dl_j}{adl} \right]} \right] \times \left[\frac{\log \left[\frac{N + 0.5}{df_s} \right]}{\log(N + 1.0)} \right]$$

tf_{ij} = frequency of the serch word i in the document j
df_s = the number of documents which contain at least one search word of the set S
S = a set of search words within the *syn* operator
dl_j = the length of the document j
adl = average document length in the collection
N = the number of documents in the collection

The *syn* operator treats all alternative translations for a search word, as expressed in a translation dictionary, as instances of the same word. For example if the Swedish word *möte* has the translation alternatives *meeting, encounter, date, appointment* in English, they are all treated as a single facet, getting the same weight as if each instance of meeting, encounter, date, appointment was replaced by the same key, e.g., meeting, before indexing.

The *od* (ordered distance) and the *uw* (unordered window) operators are proximity operators used to express target language phrases. A compound in the source language would be expressed as a phrase in the target language in the case where a direct translation in a translation dictionary is a phrase, or in the case when the compound is decomposed, the component translations are grouped as a phrase. In the tests in this thesis both the proximity operators are tested. The *od* operator (*#odN*) requires that all words in the phrase construction occur in the given order and that any adjacent words occur in a distance of less than N words.

For example, the phrase #od2(information retrieval) requires that information and retrieval occur in a distance of less than 2 words and in this order. The unordered window operator: (#uwN) requires that all words in the phrase construction must be found in any order within a window of N words.

The text database for the empirical tests in this thesis consists of English newspaper articles from the Los Angeles Times. The articles were published in 1994 and 1995 and the data was provided by CLEF for their evaluation campaign.

The indexing of the document files in the target language English is based on word form normalisation using the morphological analysis program ENGTWOL by Lingsoft plc. Finland. The analysis program produces three basic cases. First, the input word was recognised by the analyser, thus (one) basic form was produced into 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 recognised words. The third case consists of input words not recognised by the analyser (e.g. basetsane). Such words were indexed as such, preceded by a special symbol (@basetsane), thus constituting the index of unrecognised words. (Hedlund et al. 2003)

7 Evaluation methods and test environment for cross-language information retrieval

7.1 Evaluation measures

In information retrieval experiments and system evaluation the retrieval effectiveness is usually measured by the two variables *recall* and *precision.* Recall and precision are defined as:

$$\text{recall} = \frac{\text{number of relevant documents (or items) retrieved}}{\text{number relevant documents (or items) in the collection}}$$

$$\text{precision} = \frac{\text{number of relevant documents (or items) retrieved}}{\text{total number of documents (or items) retrieved}}$$

Recall measures the ability of a system to present relevant documents and precision measures the ability of the system to present only relevant documents. Although the traditional measures of recall and precision are widely used especially for system evaluation, their limitations especially regarding the usefulness of the information system to the user should be recognised (Borlund 2000; Järvelin & Kekäläinen 2000).

Precision and recall are set-based measures and evaluate the quality of an unordered set of retrieved documents. To measure ranked lists, precision and recall can be considered in combinations, for example, precision can be plotted against recall after each retrieved document. *Average precision* is a single-value measure reflecting the performance over all relevant documents. Precision can also be measured at standard recall levels (0 to 1 in increments of 0.1)

In the following the evaluation techniques and measures of the CLEF evaluation forum (Cross-Language Evaluation Forum) will be presented (Peters 2001). They are similar to those used in the "ad hoc" task of the TREC conferences (The Text Retrieval Conference). Detailed reports of evaluation techniques and measures can be found in Schäuble and Sheridan (1997), Braschler, Peters and Schäuble (1999), Voorhees (1998) and Sormunen (2002).

The CLEF evaluation results, using the *trec_eval* software, are mainly reported per each submitted run, where a run contains the results of a number of queries, but also average precision results for individual queries can be obtained. The main reported measures are: 1) Average precision (non-interpolated) for all relevant documents, 2) Interpolated Recall-Precision averages 3) A precision table reporting precision at 9 document cut-off values and 4) R-precision (precision after R documents are retrieved, when R is the number of relevant documents for the query). The result reporting in the studies included in this thesis follows the CLEF evaluation standard and has been reported on the CLEF evaluation forum.

Average precision (non-interpolated) for all relevant documents is a measure used in the studies to compare the test runs evaluating the whole query translation system and its components. It is a single-value measure and can measure the result for a whole run, containing a topic set of 40-50 topics, or it can be used to measure individual queries. Average precision is the average of the precision value obtained at each relevant document is retrieved. It rewards systems that rank relevant documents high. An example from Peters (2001) exemplifying average precision is the following: a query with four relevant documents, retrieved at ranks 1, 2, 4 and 7. The actual precision after each relevant document is retrieved is 1, 1, 0.75 and 0.57. Computing the mean, which is 0.83, gives the average precision over all relevant documents for this query as 0.83.

Interpolated recall - precision average tables were used as input for graphical presentations of the runs, where a run consists of a number of topics. Precision averages at 11 standard recall levels (0 to 1 in increments of 0.1) are used to compare system performance (see Table 1). The recall - precision average is computed by summing the interpolated precision values of the whole run at the specified standard recall cut-off value (denoted by $\sum P\lambda$ where $P\lambda$ is the interpolated precision at recall level λ) and then dividing by the number of topics

$$\frac{\sum_{i=1}^{NUM} P\lambda}{NUM} \qquad \lambda \in \{0.0, 0.1, 0.2, 0.3, ..., 1.0\}$$

Interpolated precision at standard recall level λ follows the rule to use maximum precision obtained for a query for any actual recall level $\geq \lambda$. Following this rule, precision, although not defined at a recall level of 0.0, gets an interpolated value. (Peters 2001)

Table 1 Interpolated Recall - Precision Averages

```
Total number of documents over all queries
    Retrieved:    47000
    Relevant:       856
    Rel_ret:        684
Interpolated Recall - Precision Averages:
    at 0.00       0.5861
    at 0.10       0.5109
    at 0.20       0.4661
    at 0.30       0.4128
    at 0.40       0.3648
    at 0.50       0.3294
    at 0.60       0.2676
    at 0.70       0.2344
    at 0.80       0.2089
    at 0.90       0.1739
    at 1.00       0.1319
Average precision (non-interpolated) for all rel docs(averaged
over queries)
                  0.3241
```

The precision table, reports precision at 9 document cut-off values (at 5, 10, 15, 20, 30, 100, 200, 500, 1000 retrieved documents). The precision, computed after a certain number of documents are retrieved, reflects the user's view of system performance. A user might want to look at only 30 documents instead of looking at all retrieved documents (up to 1000 or more). For each document cut-off-value, the precision average is computed by summing the precision values of each query at the specified document cut-off value (for example 30) and dividing by the number of queries if a whole test run is evaluated. (Peters 2001)

7.2 Topic creation and evaluation for cross-language information retrieval

The move of the Cross-lingual track from TREC to the CLEF evaluation forum for European languages was based on the assumption that more Europeans would join such an activity. The hope was that the knowledge of non-English languages would increase (Harman et al 2001). More non-English languages also meant new document collections and the series of topics had to be translated to many new languages. The interest of Swedish and Finnish research groups led to the inclusion of both Finnish and Swedish as topic languages already the first CLEF year and therefore could be used as source languages by the participants.

Topic creation as well as relevance assessments of documents are distributed in CLEF. The document collections, newspaper material, are from roughly the same time period, 1994-95. The topics are created on the basis of local language and cultural background which has an impact on the choice of subjects, the formation of phrases and names included in the topics. But this distributed process where topics are created in six or seven countries, originally in the native language of the creators, also makes the topic translations to the other topic languages very important. The balance between precision in translation and the naturalness with respect to the language is difficult to obtain. According to the directives for topic creation and translation for CLEF, ideally a translation should reflect how a native speaker would phrase a search for a topic in their language and culture.

The assessment of relevant documents for the topics is also a distributed process. Good assessments require a good understanding of the topic. Local assessors judge documents in their native language. However, the topics, most of them translated from other languages, can be troublesome because of their multilingual and multicultural characteristics (Harman et al. 2001).

For the recall base, the CLEF forum uses the same method as TREC, the pooling method, which means that only top-ranked documents from the participating runs are submitted to a pool, where duplicates are removed and the pool is judged by human assessors. There is no possible way to judge the relevance of all retrieved documents in the runs of the participants, however, the pooling method is considered sufficient by Voorhees (1998) and Zobel (1998).

Both workshop series for cross-language information retrieval, CLEF and NTCIR, have a common origin in TREC, but there are also differences, except for the languages covered, also in the way of performing relevance assessments. The CLEF organisation employs a binary system (like the one in TREC, where a document is either relevant or non-relevant) while NTCIR employs multi-grade relevance assessments (Kando 2001). Multi-grade relevance assessments have clear benefits (Järvelin & Kekäläinen 2000; Sormunen 2002) but are more expensive to produce.

Lists of the topics, in the languages Swedish, Finnish and German that are used in the studies in this thesis are provided in the Appendix of part I. Each complete topic includes three fields: title, description and narrative. For the test experiments included in this thesis, the title and description fields of the topics were used. The number of topics for which the CLEF organisers provided relevance assessments are for CLEF 2000, 33 topics and for CLEF 2001, 47 topics.

8 Summary of the studies

In this section a summary of the empirical studies of this thesis is presented. The research problems, the methods and the main results will be briefly summarised. Section 8.1 considers the studies on language features relevant in cross-language as well as monolingual information retrieval, with special focus on the Swedish language. Section 8.2 contributes to the area of cross-language retrieval by developing and evaluating a new system for cross-language retrieval.

8.1 Summary of Part II

The studies in this section focus on language features affecting information retrieval. There have so far been hardly any research results presented on the Scandinavian languages Swedish, Norwegian and Danish from the information retrieval point of view. Several languages, belonging to the larger group of Germanic languages possess specific features for example in compound formation, that make proper compound handling essential in cross-language information retrieval. Monolingual and cross-language information retrieval systems do not necessarily perform in a proper manner unless language specific features are recognised in the development process.

8.1.1 Article 1

The first study:

> Hedlund, T., Pirkola, A., & Järvelin, K. (2001) Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information processing & management vol. 37(1) 147-161,

analyses Swedish as a document and query language for information retrieval. Swedish is spoken as a native language by eight to nine million people, mainly in Sweden but also by a minority population in Finland. However, due to the close relationships between the Nordic countries and the other Scandinavian languages the number of people who speak Swedish and can understand it is much larger. Swedish and the other Scandinavian languages have unique and novel features that affect information retrieval, but so far very few analyses have been done.

Research problems

1) What are the features in the Swedish language relevant to information retrieval on morphologic, syntactic as well as semantic level?

2) What are the effects from the viewpoint of full text information retrieval, database indexing, query formulation and cross-language information retrieval of the
   a) morphological features, such as inflection, derivation, gender
   b) semantic features like homonymy, polysemy and hyponymy.
3) What are the pitfalls with publicly available tools for normalisation and compound splitting?

Research methods and settings

The identification of relevant linguistic features from the information retrieval point of view was based on a literature study on linguistics and information retrieval. The linguistic properties of the Swedish language were analysed and relevant features from information retrieval point of view were identified. The key concepts and an introduction to linguistics and natural language tools for information retrieval have been presented above in Sections 2 and 3 as follows:

- morphological features (Section 2.1)
- semantic features (Section 2.3)
- stemming tools (Section 3.1)
- normalisation tools (Section 3.2)

For the research, the following analysis methods were used:

- Qualitative analysis, based on literature and dictionaries of Swedish language features.
- Qualitative analysis is also performed on tools (SWETWOL) for morphological analysis. The analysis is based on selected cases.
- Statistical analysis of ambiguity based on dictionaries in three languages.
- Statistical analysis of translation ambiguity based on translation dictionaries for three language pairs.

For the study, publicly available versions of the morphological analyser SWETWOL, and off-the-shelf dictionaries for Swedish-English and Finnish-English were used.

A set of TREC topics, available in Finnish and English were also translated into Swedish. The topics contain the fields; title, description and narrative, and all words except stop words were used in the study. The topics are included in the Appendix.

Results

The description of the properties of the Swedish language points out that Swedish has a fairly rich morphology, specific features in the formation of compounds (fogemorphemes) and a high frequency of homographic words.

The fairly rich morphology indicates that stemming might not be sufficient and in the case of cross-language retrieval normalisation to base form is essential. The results of the comparative study on the degree of lexical ambiguity (homonymy and polysemy) suggest that part-of-speech tagging might be useful in Swedish due to the high frequency of homographic words.

The publicly available tools have pitfalls especially concerning compound splitting that might decrease their effectiveness in information retrieval and cross-language retrieval. The compound components might contain joining morphemes (fogemorphemes) in which case all components are not normalised to base form, e.g., *världshandel* (world trade) is decomposed into *världs* and *handel*, while the base form for *världs* is *värld* . Thus it is impossible to directly match components to the entry of a translation dictionary.

8.1.2 Article 2

The second study:

> Hedlund, T. 2002. Compounds in dictionary-based cross-language information retrieval, Information Research, 7(2) http://InformationR.net/ir/7-2/paper128.html,

concentrates on the impact of compound processing on cross-language retrieval.

Compound words form an important part of natural language as they are often content bearing words in a sentence and therefore important from the retrieval point of view. In Swedish, German and Finnish, all three compound-rich languages, around one tenth of the words in newspaper text are compounds.

The number of compounds in the test topics used in the tests is higher, about 23% in the Swedish, 25% in the German and 17% in the Finnish topics. Since only the title and description fields were used in the tests, the calculations for compounds in the topics are limited to these fields.

The topics also very often contain repetitions of compounds, naturally due to the fact that the description in many cases repeats the title. In the Swedish topics around 40% of the compounds are repeated at least once in a topic. In German the percentage is 32% and in Finnish as high as 45%.

Research questions

Important questions concerning compound handling in dictionary-based cross-language information retrieval are:

1. Should compounds be decomposed if a direct translation is not possible? That is, does compound splitting increase retrieval performance?
2. What is the effect of the normalisation of components?
3. What is the effect of dictionary-based simplistic translation of compound components, compared to an optimal translation?
4. What is the effect of phrase structuring for compound components in the target language?

Research methods and settings

A number of compound handling strategies regarding phrases, window size and word order, dictionary contents and translation methods is devised. Their performance as a part of the UTACLIR system (held otherwise stable) is evaluated using the CLEF test settings (databases, topics, relevance assessments and evaluation methods).

The process used is the UTACLIR process, an automated method for query construction for dictionary-based cross-language information retrieval. The process development and evaluation will be presented in detail in the studies in Part 3 of this thesis. The compound handling process is implemented as a component into UTACLIR. This makes it possible to evaluate the final process as well as individual steps in the process using conventional evaluation measures.

The introduction to linguistic concepts and description of the analysis tools are provided above in Sections 2, 3, 4, 5, 6 and 7 as follows:

- compounds (Section 2.2)
- phrases (Section 2.2)
- normalisation (Section 3.2)
- phrase based query structuring (Section 4.1)
- the UTACLIR process (Section 5)
- InQuery retrieval system (Section 6)
- CLEF test settings (Section 7)

Research results

The solutions proposed in this study rely on the morphological and syntactic structure of compounds. For normalisation of compound components, the features of joining morphemes in many Germanic languages, as well as the inflected components or components in the form of word stems have to be taken into account. The left- or right-branching structure for the compound components form the linguistic basis for the novel *grouping strategy* that combines and translates consecutive components in pairs.

The evaluation results of this study indicate that compound processing as a whole has a clearly positive effect on retrieval results. Since compounds are so frequent in the topics the result for the manual disambiguation was surprisingly low and the compound processing reached comparable performance. However, the test sample is relatively small, and test situations do not completely cover the complexity of natural language information retrieval. Additional test sets with different topics and different document databases could yield interesting results.

8.2 Summary of Part III.

Part III consists of three studies on the theme of developing and evaluating the UTACLIR system for query translation and construction in cross-language information retrieval. The first two studies have been presented at the Cross-language Evaluation Forum in 2000 and 2001. The third study was published in the conference proceedings of the ProLissa conference in October 2002.

8.2.1 Articles 1 and 2

The first study in Part III is:

Hedlund T, Keskustalo H, Pirkola A, Sepponen M and Järvelin K (2001) 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, Revised Papers. Lecture Notes in Computer Science 2069, Springer-Verlag, Berlin, 2001. pp. 211-225.

The second study is:

Hedlund T, Keskustalo H, Pirkola A., Airio E, and Järvelin K (2002) 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.

The studies present two versions of the UTACLIR system, and its main components, compound handling, query structure, and matching of proper names.

Research questions in study 1

1. By what process, using bilingual dictionaries, can we automatically construct effective target language queries from source language request sentences?
2. How does retrieval effectiveness vary when source languages vary?
3. How does query structure affect cross-language retrieval effectiveness when using different source languages?

Research questions in study 2

4. How do the new features for matching compound words affect retrieval effectiveness when using comprehensive translation dictionaries or a limited dictionary where all direct translation of compounds is excluded.
5. How does the new n-gram based technique for matching proper names and other non-translatable words affect retrieval effectiveness?

The new features are:

- A new process for dictionary look-up and translation of compound words.
- A new process for matching proper names and other non-translatable words.
- New ways of using stop word lists.
- Normalisation of dictionary output.

Research methods and settings
Both studies involve development and testing of the UTACLIR query translation system.

The evaluation of system performance is done according to the rules and practice of the TREC and CLEF evaluation forums. This means that the test topics, document data, and relevance assessments were provided by the CLEF evaluation forum. For all runs, queries were constructed on the basis of the title and description field of the topics. The InQuery system was used as the retrieval software. Average precision for all the queries in each run is used as the evaluation measure of system performance. Additionally, average precision for each document cut-of-value and interpolated recall-precision average tables are reported.

A description of the concepts and tools is provided as follows:

- the UTACLIR process (Section 5)
- CLEF evaluation procedure and test environment (Section 7)
- the InQuery system (Section 6)
- CLEF topics (Appendix)
- structuring of queries (Section 4.1)
- handling of untranslatable words (Section 4.1)

Study 1 consists of four official test runs (Finnish - English structured queries, Swedish - English structured queries, German - English structured queries and German - English unstructured queries) and two additional test runs (Finnish - English unstructured queries and Swedish - English unstructured queries).

Study 2 consists of four official test runs (four automated bilingual runs, three language pairs, Finnish – English, Swedish – English and German – English). As additional runs we have performed test runs eliminating the n-gram algorithm from the process.

Research results
Study 1.

The automated dictionary-based UTACLIR process in its first version proved its effectiveness in translating and constructing queries from source language request sentences to the target language. Thus the UTACLIR framework, focusing on language specific features in the source and target language, such as morphology and compounds, forms an important basis for further developments of the process.

The official test results of the runs show comparable performance for three different source languages using the UTACLIR query translation method. In fact the retrieval effectiveness was surprisingly stable when changing the source language.

The structured / unstructured query performance for all language pairs indicates better performance for structured queries. In this case structuring of queries was most successful for Finnish. For Swedish and German the effect was smaller but also in these cases positive and clear.

Study 2.

Generally, the results for all the four runs were good. Average precision for all the queries shows clear improvements, but there still is great variation in the performance of individual queries. Some queries perform exceedingly well getting high scores, but some fail to retrieve relevant documents. This holds for all language pairs.

The effect of removing compounds from the dictionary – simulating a small dictionary was a test in order to establish the effectiveness of the compound handling process. For German – English we 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 fields) 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 drop in average precision for the run with the limited dictionary is relatively small, from 0.3474 to 0.3054 or 4.2 % units.

The results for the four unofficial runs testing the effect of n-grams indicate that eliminating the n-gram algorithm from the automated process results in a decline in performance (average precision value for all the 47 queries) for the Finnish – English and the Swedish – English runs. On the other hand, both German runs perform better without the n-gram algorithm. For individual queries the result varies. The reason for the decline in the result seems to be that noise is added to the queries by the use of the n-gram algorithm. Even though the algorithm matches the right index words, as in for example the proper name “Euskirchen” and the abbreviation “rsi”, very common words as “kitchen" and "crisis” are added to the queries. On the other hand, the simplistic approach of passing German source query proper names as such to the English target query is often successful enough.

8.2.3 Article 3

The third study of Part III is:

Hedlund,T., Pirkola, A, Keskustalo, H., Airio, E. and Järvelin K. (2002) Cross-language information retrieval: using multiple language pairs. In Bothma T., Kaniki A. eds. Progress in Library and Information Science in Southern Africa. Proceedings of the second biennial DISSAnet Conference. 24-25 October 2002, Farm Inn, Pretoria, South Africa.

Based on this conference paper the article:

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K. (2003) Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002

appears in the international journal Information Retrieval.

In this study we present performance analysis results for the automated dictionary-based cross-language UTACLIR query translation software. The process is tested in two developmental stages and the focus is to test the robustness of the process with a large topic set. The tests involve three language pairs (Swedish - English, Finnish - English and German - English).

Research questions

Our interest is in evaluating the whole process as well as the contribution of its components. The aim is to get a broad and more complete view of the robustness of a process in development by using a large test topic set in three source languages. We wanted to know:

- What is the contribution of the components, the compound handling process, the n-gram based matching of proper names and finally the structuring of the target query.

Research methods and settings

The evaluation results for the whole process as well as the components have been established through several tests with a large topic set of 80 topics provided by the CLEF evaluation forum in the years 2000 and 2001. The test topics are in three languages, Finnish, Swedish and German plus a topic set in English used for a monolingual baseline run for comparison. The document collection and the appropriate relevance files are also provided by CLEF. The document database in English contains newspaper articles published in the Los Angeles Times.

The performance of the individual components; the component for handling compound words and the component for handling untranslated words are evaluated as part of the UTACLIR system (held otherwise stable).

Research results

The contributions of this study are:

- Performance analysis results of a dictionary-based CLIR system for three 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 and query structuring.

In addition to the large number of topics, the sets are in three languages, Swedish, Finnish and German. The process is tested in two development stages. In general the test results indicate that the process is quite robust and transferable 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 from CLEF 2000 are lower in each of the runs regardless of topic language. This is also true for the monolingual run. There is a drop in average precision of the cross-language queries to the monolingual baseline, which of course is to be expected.

The automated CLIR processes are also tested for their component performance. The tests with structuring 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 is particularly noticeable for Finnish and German where the translation dictionaries are comprehensive.

The compound handling process for compound rich languages is important. Splitting compounds into constituents enables translation of the individual constituents that often are content bearing words. In this study where the target language is English, settings allowing a phrase-based structure of compound components 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 a synonym structure the results were better. The results are consistent with what has been established in earlier studies on English monolingual retrieval (Mitra, Buckley, Singhal & Cardie 1997). 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, since phrase structure also has a clear disambiguation effect in CLIR.

The n-gram algorithm was implemented in the process in order to solve the problem of untranslated words, such as proper names. The process was particularly successful for the Finnish language where proper names usually appear in inflected forms and where matching to the document index therefore is difficult. However, the n-gram algorithm in this setting identified a set of six most similar words - three most similar recognised by the morphological analyser and three unrecognised words. This selection was not probably ideal for German where in several cases the untranslated word appeared in an identical form in the source and target language and no n-gram matching would have been needed. In these cases noise was added to the queries by the algorithm, which resulted in a decrease in performance for German.

The results indicate that, in all, the process is robust and can be transferred to different languages with small differences in the performance. The individual effects of the different components are in general positive, however, the performance also depends on the topic set, the number of compounds and proper names, and to some extent on the source language used.

9 Discussion and conclusion

The main contributions of this study are summed up as follows:

- The study of cross-language information retrieval problems of using Swedish morphological analysis tools. The problems concern compounds with joining morphemes as well as the high frequency of homographic words in Swedish.
- The design of a dictionary-based method for cross-language information retrieval, the UTACLIR method for the automated processing of query translations for multiple language pairs. The method involves solutions for handling compounds and problem words.
- Empirical testing of the UTACLIR system as a whole and of its individual components over a large topic set.

When the research work on this thesis started, there was very limited knowledge on information retrieval in Swedish. Neither had there been any cross-language information retrieval research in Swedish. Therefore serious research attention was in order. This was based on the fact that 1) the Nordic countries have a fairly large population, 2) they are technically advanced countries with lots of activity related to information retrieval, and because 3) the features of Swedish differ from those of many other languages, such as English, French, Spanish and Finnish. In the first study properties of the Swedish language relevant from information retrieval viewpoint are described, and a number of research problems for Swedish information retrieval are pointed out. The problems with compound words and joining morphemes are followed up in the later CLIR-studies. Similar problems have been reported for German text retrieval in Braschler & Ripplinger (2003) and in Ripplinger (2001).

The importance of compound processing in dictionary-based cross-language information retrieval is shown. A finding of the study on compounds was that around ten percent of content words in running text are compounds in the three source languages studied (Swedish, Finnish and German). This can be reformulated even more impressively: it means that more than twenty percent of morphemes in running text are within compounds. The compound handling process was implemented as a component in the automated cross-language information retrieval system UTACLIR. The solutions proposed in this study rely on the morphological and syntactic structure of compounds. For normalisation of compound components, the features of fogemorphemes in Germanic languages, as well as the inflected components or components in the form of word stems have to be taken into account.

The evaluation results of this study indicate that compound processing as a whole has a clearly positive effect on retrieval results. The relative importance of compound processing in dictionary based cross-language information retrieval naturally depends on the number of compounds present in source language requests as well as on the available translation resources. In the German, Swedish and Finnish CLEF topics over 20% of the remaining words after stop word removal are compound components. A direct translation of compounds is clearly more precise and therefore effective. Using comprehensive dictionaries one may directly translate more compounds than with limited and small translation dictionaries. However, on the other hand, no dictionary can hold entries for all occasional compounds in a language. The test with two dictionaries for German as the source language gives an indication that the compound splitting features in the UTACLIR process work well also with a limited dictionary.

The analysis of compounds and their features from the cross-language information retrieval point of view in this study could be extended to semantic analysis to determine, for example, paradigmatic relations and to syntactic analysis to determine syntactic structures of phrases. This kind of information could be used in information retrieval and cross-language information retrieval applications to determine valuable keys.

A way to handle the random effects of chance in scientific studies is to use large samples. In information retrieval using large topic sets in the experiments can eliminate the effects caused by chance. In addition to the large number of topics, the sets used in the concluding study are in three languages, Swedish, Finnish and German, plus a topic set in English used for a monolingual run for comparison. The process was also tested in two development stages. The automated CLIR processes were also tested for their component performance. The tests with queries structured by the Pirkola method confirm 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 is particularly noticeable for Finnish and German where the translation dictionaries were comprehensive.

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

The findings in this study for the bilingual processes indicate that a similar process could also be used in a multilingual environment. The basic structure of the process is transferable to other languages, and the language-specific features in the component processes can be adapted. The results of the study also indicate a positive outcome for the combination of linguistic tools and a probabilistic matching technique and structuring of queries. The linguistic approaches using normalisation, and compound handling as well as all the other components, have a considerable value-adding effect, but the optimisation of the combined effect remains a challenge for further research.

References

Akmajian, A., Demers, R., Farmer, A., Harnish, R. (1995). *Linguistics: An introduction to language and communication*, 4th ed. MA: the MIT press.

Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval* 4, pp. 95-208.

Allan, J., Callan, J., Croft, B., Ballesteros, L., Broglio, J, Xu, J. & Shu, H. (1997) INQYERY at TREC 5. In *The Fifth Text Retrieval Conference (TREC 5)*
[http://trec.nist.gov/pubs/trec5/t5_proceedings.html Accessed 02.10.2003]

Ballesteros, L. & Croft, W.B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, PA, USA , pp. 84-91.

Ballesteros, L. & Croft, W.B. (1998a). Statistical methods for cross-language information retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval, 23-40.* Boston: Kluwer Academic Publishers.

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

Bauer, L. (1983). *English word formation.* Cambridge: Cambridge University Press.

Blåberg, O. (1988). *A study of Swedish compounds.* Department of General Linguistics, University of Umeå, Report 29.

Borlund, P. (2000). *Evaluation of interactive information retrieval systems*. Diss. Åbo Akademi University. Åbo: Åbo Akademi University Press.

Braschler, M. (2001). CLEF 2000 - Overview of results. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069.* Berlin: Springer.

Braschler, M. & Ripplinger, B. (2003). Stemming and Decompounding for German Text Retrieval In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003. Lecture Notes in Computer Science 2633.* Berlin: Springer.

Braschler, M., Peters, C. & Schäuble, P. (1999). Cross-language information retrieval (CLIR) track overview. In *Proceedings of the eight Text Retrieval Conference 8(TREC8).* [http://trec.nist.gov/pubs/trec8/t8_proceedings.html Accessed 02.10.03]

Broglio, J., Callan, J. & Croft, W.B. (1994). Inquery system overview. In *Proceedings of the TIPSTER Text Program (Phase I),* pp. 47-67.

Buckley, Singhal, Mitra & Salton (1996). New retrieval approaches using SMART: TREC-4. In *Proceedings of the fourth Text Retrieval conference (TREC4).* [http://trec.nist.gov/pubs/trec4/t4_proceedings.html Accessed 02.10.03]

Carlberger, J., Dalianis, H., Hassel, M. & Knutsson, O. (2001). Improving Precision in Information Retrieval for Swedish using Stemming. *In the Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001,* Uppsala, Sweden.

Croft, W.B., Turtle, H.R. & Lewis, D.D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,Chicago, Illinois, USA*, pp. 32-45.

Davis, M. (1998) On the effective use of large parallel corpora in cross-language text retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval,* pp. 11-22. Boston: Kluwer Academic Publishers.

Davis, M. & Ogden, W. C. (1997). QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System. In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, PA, USA, pp. 92-98.

Fleischer,W. & Barz, I. (1992). *Wortbildung der deutschen Gegenwartssprache*. Tübingen: Max Niemeyer Verlag.

Fluhr, C., Schmit, D., Ortet, P., Elkateb, F., Gurtner, K. & Radwan, K. (1998). Distributed cross-lingual information retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval, 41-50.* Boston: Kluwer Academic Publishers.

Fox, C. (1990). A stop list for general text. *ACM SIGIR Forum* 24, 19-35.

Gachot, D., Lange, E. & Yang, J. (1998). In G. Grefenstette ed. *Cross-Language Information Retrieval, 105-118.* Boston: Kluwer Academic Publishers.

Gey, F., Jiang, H., Petras, V. & Chen, A. (2001) Cross-language retrieval for the CLEF collections - comparing multiple methods of retrieval. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069.* Berlin: Springer.

Gollins, T. & Sanderson, M. (2001) Sheffield University CLEF 2000 submission - bilingual track: German to English. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069.* Berlin: Springer.

Gonzalo, J. (2001). Language resources in cross-language text retrieval: A CLEF perspective. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069.* Berlin: Springer.

Grefenstette, G. (1998). The problem of Cross-Language Information Retrieval. In G. Grefenstette (Ed.) *Cross-Language Information Retrieval, 1-9.* Boston: Kluwer Academic Publishers.

Haas, S. (1996) Natural language processing: towards large-scale, robust systems. In *Annual Review of Information Science and Technology*, 31. Medford, New Jersey: American Society for Information Science, pp. 83-119.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), pp. 7-15.

Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P. & Sheridan, P. (2001). CLIR Evaluation at TREC. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069.* Berlin: Springer.

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 structuring. In Peters, C. ed. *Cross-language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069,* Berlin: Springer

Hedlund T, Keskustalo H, Airio E & Pirkola A (2002). UTACLIR - an Extendable Query Translation System. Paper presented at the *ACM SIGIR Workshop for Cross-Language Information Retrieval, August 15th 2002* in Tampere, Finland.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. (2003) Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. Manuscript, accepted for publication in *Information Retrieval*.

Hiemstra, D., Kraaij, W., Pohlmann, R. & Westerveld, T. (2001) Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers.* Lecture Notes in Computer Science 2069. Berlin: Springer.

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

Hull, D. (1998). A weighted Boolean model for cross-language text retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval, 119-136.* Boston: Kluwer Academic Publishers.

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

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Atens, Greece pp. 41-48.

Kando, N. (2001) NTCIR Workshop: Japanese- and Chinese-English Cross-Lingual Information and Multi-grade Relevance Judgements. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers.* Lecture Notes in Computer Science 2069. Berlin: Springer.

Karlgren, J. (2000) *Information retrieval: Statistics and linguistics.* Stockholm: Swedish Institute for Computer Science. [http://www.sics.se/~jussi/Undervisning/texter/ir-textbook.pdf Accessed 28.05.2003]

Karlsson, F. (1994). *Yleinen kielitiede*. [General linguistics] Helsinki: Yliopistopaino [In Finnish].

Karlsson, F. (1992). SWETWOL: A comprehensive morphological analyser for Swedish *Nordic Journal of Linguistics,* 15, pp. 1-45.

Kekäläinen, J., & Järvelin, K. (1998) The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, August 24-28th 1998 pp. 130-137.

Koskenniemi, K. (1983) *Two-level morphology. A general computational model of word-form recognition and production.* Publications of the Department of General Linguistics, University of Helsinki. No. 11.

Kraaij, W. & Pohlman, R. (1996). Viewing stemming as recall enhancement. In *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Zürich, Sweitzerland, pp. 40-48.

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

Landauer, T.K. & Littman, L.M. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the 6th Conference of University of Waterloo Centre for the New Oxford English Dictionary and Text Research,* pp. 31-38, Waterloo, March 1990.

Levi, J. (1978). *The syntax and semantics of complex nominals.* New York: Academic Press.

Littman, M., Dumais, S. & Landauer, T. (1998). Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette ed. *Cross-Language Information Retrieval, 51-62.* Boston: Kluwer Academic Publishers.

Lovins, J. B. (1968). Development of a stemming algorithm. *Translation and computational linguistics.* 11(1), pp. 22-31.

Lyons, J. (1981). *Language and linguistics: An introduction*. Cambridge: Cambridge University Press.

Malmgren, S. G. (1994). *Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser.* Lund: Studentlitteratur [Swedish lexicology. Words, word formation, dictionaries and word databases.]

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery* 7, 216-244. Reprinted in K. Sparck Jones, P. Willet, Eds. *Readings in Information Retrieval.* San Francisco, CA: Morgan Kaufmann 1997.

McNamee, P., Mayfield, J. & Piatko, C. (2001). A language-independent approach to European text retrieval. In Peters, C. ed. *Cross-language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop, Lecture Notes in Computer Science 2069,* Berlin: Springer

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

Moulinier, I., McCulloh, A. & Lund, E. (2001). West group at CLEF 2000: Non-English monolingual retrieval. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers.* Lecture Notes in Computer Science 2069. Berlin: Springer.

Noreen, A. (1903 - 1907). *Vårt språk V.* Lund: C.W.K. Gleerups förlag.

Oard, D. (1997). Alternative approaches for cross-language text retrieval. Paper presented at the AAAI Spring Symposium on Cross Language Text and Speech Retrieval, Palo Alto CA, March 1997. [http://citeseer.nj.nec.com/oard97alternative.html Accessed 02.10.03]

Oard, D. & Dorr, B. (1996). A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19. University of Maryland, Institute for Advanced Computer Studies. [http://citeseer.nj.nec.com/oard96survey.html Accessed 02.10.03]

Oard, D., Levow, G-A. & Cabezas, C. (2001). CLEF experiments at Maryland: Statistical stemming and backoff translation. In C. Peters (Ed.) Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069. Berlin: Springer.

Peters, C. ed. (2001). Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069. Berlin: Springer.

Pfeifer, C., Poersch, T. & Fuhr, N. (1996). Retrieval effectiveness of proper name search methods. Information Processing and Management, 32(6), pp. 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*, Melbourne, Australia, August 24-28th 1998. pp. 55-63.

Pirkola, A. (1999) *Studies on linguistic problems and methods in text retrieval: The effects of anaphor and ellipsis resolution in proximity searching and translation and query structuring methods in cross-language retrieval.* Ph.D. Thesis, University of Tampere, Department of Information Studies, Acta Universitatis Tamperensis 672.

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

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, 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/paper126.html Accessed 28.05. 2003.]

Pirkola, A., Puolamäki, D. & Järvelin, K. (2003). Applying query structuring in cross-language retrieval. *Information Processing and Management* 39, pp. 391-402.

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

Ripplinger, B. (2001). The use of NLP techniques in CLIR. In C. Peters (Ed.) *Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers.* Lecture Notes in Computer Science 2069. Berlin: Springer.

Robertson, S. E. (1977). The probability ranking principle in IR *Journal of Documentation,* 33, 294-304. Reprinted in K. Sparck Jones, P. Willet, Eds. *Readings in Information Retrieval.* San Francisco, CA: Morgan Kaufmann 1997.

Salton, G. (1970). Automatic text processing of foreign language documents. *Journal of the American Society for Information Science,* 21, pp. 187-194.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM* 18, 613-620. Reprinted in K. Sparck Jones, P. Willet, Eds. *Readings in Information Retrieval.* San Francisco, CA: Morgan Kaufmann 1997.

Salton, G. & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

SAOL (1986). *Svenska Akademiens ordlista.* 11 upplagan. Stockholm: Norstedts [Word list of the Swedish Academy].

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), pp. 944-952.

Savoy, J. (2002). Report on CLEF-2001 experiments. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck, Eds. *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001.* Lecture Notes in Computer Science 2406, Berlin: Springer, 2002.

Schäuble, P. & Sheridan, P. (1997). Cross-language information retrieval (CLIR) track overview. In *Proceedings of the sixth Text Retrieval Conference (TREC6)*

Sheridan, P. & Smeaton, A.F. (1992). The application of morpho-syntactic language processing to effective phrase matching *Information Processing and Management* 28(3) pp. 349-369.

Sormunen, E. (2002). Liberal relevance criteria of TREC - Counting on negligible documents? In *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval.* Tampere, Finland, pp. 324-330.

Sparck Jones, K. (1983). So what about parsing compound nouns? In Karen Sparck Jones, Yorik Wilks, Eds. *Automatic Natural Language Parsing.* New York: Ellis Horwood pp. 164-168.

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

Strzalkowski, T. (1996). Natural language information retrieval. *Information Processing and Management* 31(3), pp. 397-417.

Strzalkowski, T. (Ed.) (1999). *Natural language information retrieval*. Dordrecht: Kluwer

Strzalkowski, T., Lin, F., Wang, J. & Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. In T. Strzalkowski, ed. *Natural language information retrieval*. Dordrecht: Kluwer

Teleman, V., Hellberg, S., & Andersson, E. (1999). Svenska Akademiens grammatik 1-4. Stockholm: Svenska Akademien [Grammar of the Swedish Akademy 1-4]

Tomlinson, S. (2002). Stemming evaluated in 6 languages by Hummingbird SearchServer at CLEF 2001. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck, Eds. *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001.* Lecture Notes in Computer Science 2406, Berlin: Springer, 2002.

Tzoukermann, E., Klavans, J. & Jacquemin, C. (1997). Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging and shallow parsing. In *Proceedings of the 20th ACM/SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, PA, USA, pp. 148-155.

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

Voorhees, E. (1998). Variations in relevance judgements and the measurement of retrieval effectiveness, *Information Processing & Management,* 36, pp. 697-716.

Vossen, P. (1997). EuroWordNet: A multilingual database for information retrieval. In *Third DELOS Workshop Cross-Language Information Retrieval*. pp. 85-94. European Research Consortium for Informatics and Mathematics.

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

Yamabana, K., Muraki, K., Doi, S. & Kamei, S. (1998). A language conversion front-end for cross-language information retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval, 93-104.* Boston: Kluwer Academic Publishers

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval.* Melbourne, Australia, 307-314.

Appendices

CLEF 2000, English topics 001-040

CLEF 2001, English topics 041-090

CLEF 2000, English topics

<num> C001
<E-title>
Architecture in Berlin
<E-desc>
Find documents on architecture in Berlin.
<E-narr>
Relevant documents report, in general, on the architectural features of Berlin or, in particular, on the reconstruction of some parts of the city after the fall of the Wall.

<num> C002
<E-title>
The Electroweak Theory
<E-desc>
Find documents that report recent discoveries in the field of subnuclear physics that confirm the unified electroweak theory of Weinberg-Salam-Glashow.
<E-narr>
Relevant documents report on discoveries in the last ten years of subatomic particles, such as quarks or photons, which provide experimental confirmation of the standard theoretical model of nuclear interactions proposed by Weinberg-Salam-Glashow. Other work in the field of nuclear physics that does not have a direct connection with this theory is not pertinent.

<num> C003
<E-title>
Drugs in Holland
<E-desc>
What is the drugs policy in the Netherlands?
<E-narr>
Relevant documents report regulations and decisions made by the Dutch government regarding the sale and consumption of hard and soft drugs.

<num> C004
<E-title>
Floods in Europe
<E-desc>
Find documents that give figures on the economic costs of the damage to agriculture caused by floods in Europe.
<E-narr>
Relevant documents will report numerical data regarding damage in terms of financial losses or the loss of large quantities of goods for all sectors of

agriculture (both animals and crops) caused by flooding in Europe. Documents that discuss economic assistance to farmers to recover from flood damage are not relevant.

<num> C005
<E-title>
European Union Membership

<E-desc>
Identify attitudes of non-member countries toward
joining the European Community or European Union.

<E-narr>
Relevant documents report views expressed by a national or a spokesperson of EU
non-member
countries with respect to joining the European Union.

<num> C006
<E-title>
French Conscientious Objectors
<E-desc>
What tasks are given to French conscientious objectors during their national service?
<E-narr>
In France, conscientious objectors must do twenty months of non-military national service. Relevant documents mention the kinds of jobs or tasks that they are given during these twenty months.

<num> C007
<E-title>
Drug Use and soccer
<E-desc>
Find documents about drug use in soccer.
<E-narr>
Relevant documents report cases of soccer players convicted of taking drugs.
General discussions on drug-related issues in the world of soccer are also relevant.

<num> C008
<E-title>
The Suicide of Pierre Bérégovoy
<E-desc>
Find documents on the suicide of former French Prime Minister Pierre Bérégovoy.

<E-narr>
Pierre Bérégovoy was the French Prime Minister from 1992 to 1993. He
committed suicide two months after his sudden resignation. Relevant documents report on the circumstances and the possible reasons for his suicide, as well as the reactions that it provoked.

<num> C009
<E-title>
Methane Deposits
<E-desc>
What locations throughout the world have methane gas deposits which can be converted to energy use?
<E-narr>
Discussions of successes and efforts to locate and exploit sources of methane for commercial use are relevant. Also of interest is evidence of problems, costs, and expenditures for methane access and processing.

<num> C010
<E-title>
War and Radio
<E-desc>
What role do radios play during war or armed conflict?
<E-narr>
In warring countries, the radio plays an important role by broadcasting propaganda and messages of hope or warning. Relevant documents talk about this role, for instance during World War II or the war in Rwanda.

<num> C011
<E-title>
New Constitution for South Africa
<E-desc>
Find documents discussing the new constitution of South Africa and its final structure.
<E-narr>
Political discussions and decisions taken in preliminary negotiations between political and societal groups are relevant. The final result of the consultations and of the plebiscite for the new constitution of the Republic of South Africa is also of interest.

<num> C012
<E-title>
Solar Temple
<E-desc>
Find documents on the Order of the Solar Temple and the murder and/or
suicide of its members.
<E-narr>
The Order of the Solar Temple, a sect founded by Luc Jouret and Joseph di
Mambro, became famous after the death of nearly all its members in Switzerland and in Canada.
All documents about this sect are relevant.

<num> C013
<E-title>

Conference on Birth Control
<E-desc>
What were the discussions and resolutions at the World Population Conference on birth control in Cairo?
<E-narr>
All political discussions, proposals and resolutions on birth control at the World Population Conference are of interest. The positions of different countries, organisations and groups are especially relevant.

<num> C014
<E-title>
Tourism in the U.S.
<E-desc>
What countries, other than Canada, are sending large numbers of tourists to the U.S., and what are the principal destinations of these tourists?
<E-narr>
To be relevant, a document will indicate a country or countries which sends large numbers of tourists (e.g., more than 100,000 a year) to the U.S., and the document will further specify principal geographic areas of their interest.

<num> C015
<E-title>
Competitiveness of European Industry
<E-desc>
What are the factors that damage the competitiveness of European industry on the world's markets?
<E-narr>
Relevant documents discuss factors that render European industry and
manufactured goods less competitive with respect to the rest of the world,
e.g. North America or Asia. Relevant documents must report data for Europe as a whole rather than for single European nations.

<num> C016
<E-title>
The French Academy
<E-desc>
Find documents on the French Academy.
<E-narr>
The French Academy is an organization intended to protect and develop the French language. Relevant documents report on its influence in France and in other French speaking countries and discuss its activities. These activities include distributing awards for literature and the inclusion of new words in the
'Dictionary of the French Academy'. Documents that mention the names of some
or all of its members are not relevant.

<num> C017
<E-title>
Bush Fire near Sydney
<E-desc>
What was the extent of the bush fires near Sydney and how dangerous was the situation?
<E-narr>
Documents describe the extent and threats of bush fires near
Sydney. They also give details of the consequences of these fires.

<num> C018
<E-title>
Firefighter Casualties
<E-desc>
Find documents that discuss injuries to people engaged in firefighting.
<E-narr>
Incidences of serious injury or death to firefighters (professionals or volunteers)
around the world are relevant. Also relevant is evidence of improved measures to prevent injury.

<num> C019
<E-title>
Gulf War Syndrome
<E-desc>
Find documents that discuss the Gulf War syndrome.
<E-narr>
All references to Gulf War syndrome are considered relevant.

<num> C020
<E-title>
Single European Currency
<E-desc>
What are the advantages and disadvantages of a single European currency?
<E-narr>
Relevant documents report opinions concerning the advantages and disadvantages of the creation of a single currency for the European Union. Documents that discuss the ways in which the unification of the currency will be achieved are not relevant.

<num> C021
<E-title>
European Economic Area
<E-desc>
Find documents about the European Economic Area (EEA).
<E-narr>
The European Economic Area (EEA) is an association consisting of the members of the European Union and some other European states. Relevant documents discuss its role in the European or world economy. Documents that list one or

more members of the EEA are also relevant.

<num> C022
<E-title>
Airplane Runway Accidents
<E-desc>
Identify instances where airplane accidents have occurred on runways during flight operations.
<E-narr>
Of interest would be the causes of airplane accidents on runways and what, if anything, is being done about it. Accidents occurring by landing short of the
runway or losing power on takeoff are not considered runway accidents.

<num> C023
<E-title>
Postmenopausal Pregnancy
<E-desc>
In 1994, a postmenopausal Italian woman delivered a son, becoming the oldest
woman in the world to give birth. What was her name and how old was she?
<E-narr>
In 1994, as the result of a new hormonal therapy, a postmenopausal Italian woman delivered a son, becoming the oldest woman in the world to give birth. Relevant documents report the name of this woman and her age at time of delivery.

<num> C024
<E-title>
World Trade Organization
<E-desc>
What are the views in the U.S. in favor of GATT/ World Trade Organization?
<E-narr>
Discussions that favor U.S. membership in the World Trade Organization (WTO) are relevant. Focus is on the views that report specific benefits for the U.S. economy in general and U.S. trade in particular.

<num> C025
<E-title>
Corruption in Italy
<E-desc>
Which Italian ex-ministers have been imprisoned on the charge of accepting
bribes?
<E-narr>
Documents must contain the names of Italian ministers who have been arrested and put into prison with the accusation of bribery and corruption.

<num> C026
<E-title>

Use of Wind Power

<E-desc>

Give examples of the use of wind power

<E-narr>

Theoretical discussion or actual instances of the use of wind power are relevant. Examples of wind damage are not relevant.

<num> C027

<E-title>

Integration of German Immigrants

<E-desc>

Why do German late immigrants find integrating difficult?

<E-narr>

Relevant documents discuss the problems encountered by people of German origin from Eastern Europe coming to live in Germany. Problems regarding living conditions, education, language, employment, etc. are particularly important.

<num> C028

<E-title>

Teaching Techniques for non-English Speakers

<E-desc>

What techniques are used in U.S. schools to teach students whose native language is other than English?

<E-narr>

Relevant reports will include a discussion of the trends in education of immigrant students in the U.S., the difficulties encountered and the relative
success of various teaching methods.

<num> C029

<E-title>

Nobel Prize for Economics

<E-desc>

Who won the Nobel Prize for Economics in 1994 and for what theory?

<E-narr>

Find documents that give both the names of the winners of the Nobel prize for
Economics in 1994 and the name of the theory for which the prize was awarded.

<num> C030

<E-title>

Supermarket Ceiling in Nice collapses

<E-desc>

What caused a supermarket in Nice to cave in and what were the consequences?

<E-narr>

Find documents describing and discussing the collapse of a supermarket ceiling in Nice. All information on this accident and its causes are of interest. Discussions on its consequences are also relevant.

<num> C031
<E-title>
Consumer Protection in the EU
<E-desc>
Find reports on consumer protection in the European Union.
<E-narr>
Find information on consumer protection in the Eurpean Union (EU). Legal aspects (laws, directives, sentences of courts etc.) and political discussions are also relevant.

<num> C032
<E-title>
Female priests
<E-desc>
Find documents on the ordination of women in European Churches.
<E-narr>
Some European Churches have decided to approve the ordination of women as
priests or pastors. All documents that discuss this change of direction and reactions to it, especially in the Vatican, are of interest.

<num> C033
<E-title>
Cancer Genetics
<E-desc>
Find documents discussing recent discoveries on the relationship between
genetics and cancer.
<E-narr>
The gene responsible for the development of certain tumors has recently
been discovered. All documents that provide information on this new diagnostic
tool are relevant.

<num> C034
<E-title>
Alcohol Consumption in Europe
<E-desc>
Find documents on alcohol consumption in Europe.
<E-narr>
General information on alcohol consumption in Europe is of interest. Documents discussing alcohol abuse are also relevant.

<num> C035
<E-title>

Wolves in Italy
<E-desc>
Where can wolves be found in Italy?
<E-narr>
Wolves can be found in a number of Italian regions, in mountainous areas, in woods and in the natural parks. Relevant documents will report the name of areas or places where wolves can be found.

<num> C036
<E-title>
Olive Oil Production in the Mediterranean
<E-desc>
How much olive oil is produced in the Mediterranean area?
<E-narr>
Find information about olive oil production in the different Mediterranean countries and in the area as a whole. Only economic aspects are of interest. Documents discussing problems of quality are not relevant.

<num> C037
<E-title>
Sinking of the Estonia
<E-desc>
Find documents describing the sinking of the Estonia ferry in the Baltic Sea and reporting ongoing investigations.
<E-narr>
Information on the sinking of the Estonia ferry and the reasons for this accident is relevant. Ongoing investigations into this accident and its consequences for security in the future construction of ferryboats are also of interest.

<num> C038
<E-title>
Return of Military Remains
<E-desc>
Identify instances of the remains of deceased military personnel being returned
to their home country for reburial.
<E-narr>
In addition to the return home of any military remains, any other actions
taken or being considered to facilitate the return should be considered relevant.

<num> C039
<E-title>
Investments in Eastern Europe or Russia
<E-desc>
Identify companies that are making or have made investments in Eastern Europe
or Russia following the break-up of the former USSR.

<E-narr>
A relevant report must mention the name or owner of companies investing in Eastern Europe or Russia after the break-up of the USSR and will also give the name of the particular country or countries.

<num> C040
<E-title>
Privatisation of German Rail
<E-desc>
Find documents on the privatisation of the German railways.
<E-narr>
All aspects of the privatisation of the German railways (Deutsche Bundesbahn) are of interest. In particular, documents discussing economic aspects as well as changes in the services provided and the new management and personnel structure are relevant.

<num>C041</num>
<EN-title>Pesticides in Baby Food</EN-title>
<EN-desc>Find reports on pesticides in baby food.</EN-desc>
<EN-narr>Relevant documents give information on the discovery of pesticides in baby food. They report on different brands, supermarkets, and companies selling baby food which contains pesticides. They also discuss measures against the contamination of baby food by pesticides.

<num>C042</num>
<EN-title>U.N. Invasion of Haiti</EN-title>
<EN-desc> Find documents on the invasion of Haiti by U.N./US soldiers.</EN-desc>
<EN-narr>Documents report both on the discussion about the decision of the U.N. to send US troops into Haiti and on the invasion itself. They also discuss the direct consequences.

<num>C043</num>
<EN-title>El Niño and the Weather</EN-title>
<EN-desc>Find reports explaining the "El Niño" phenomenon and its repercussions on the world’s weather (including effects on temperature, air pressure, rain fall, etc.).</EN-desc>
<EN-narr>Relevant documents will give information on the effects of "El Niño". Interactions between the oceans and the earth's atmosphere are of interest with respect to this phenomenon. "El Niño" is especially important in the southern Pacific because of the influence it has on the world’s climate.</EN-narr>

<num>C044</num>
<EN-title>Indurain Wins Tour</EN-title>
<EN-desc>Reactions to the fourth Tour de France won by Miguel Indurain.</EN-desc>
<EN-narr>Relevant documents comment on the reactions to the fourth consecutive victory of Miguel Indurain in the Tour de France. Also relevant are documents discussing the importance of Indurain in world cycling after this victory.</EN-narr>

<num>C045</num>
<EN-title>Israel–Jordan Peace Treaty</EN-title>
<EN-desc> Find reports citing the names of the main negotiators of the Middle East peace treaty between Israel and Jordan and also documents giving detailed information on the treaty.</EN-desc>
<EN-narr>A peace treaty was signed between Israel and Jordan on 26 October 1994 opening up new possibilities for diplomatic relations between the two countries. Relevant documents will give details of the treaty and/or will name the principal people involved in the negotiations.</EN-narr>
<num>C046</num>

<EN-title>Embargo on Iraq</EN-title>
<EN-desc>What effects has the U.N. embargo had on the lives of the Iraqi people?</EN-desc>
<EN-narr>Documents describing changes in the life of the Iraqi people by comparing life before the embargo to life afterward are relevant, provided the change is directly attributable to the embargo itself. Documents containing unsubstantiated rhetoric such as "Letters to the Editor" or reports from clearly biased parties for political purposes are not relevant.</EN-narr>

<num>C047</num>
<EN-title>Russian Intervention in Chechnya</EN-title>
<EN-desc>What are the reasons for the military intervention of Russia in Chechnya?</EN-desc>
<EN-narr>Relevant documents will discuss the reasons and underlying motivations behind the intervention of Russian troops in Chechnya. Declarations by Russian politicians, including President Yeltsin, that justify the sending of Russian troops to Cechnya will also be considered pertinent.</EN-narr>

<num>C048</num>
<EN-title>Peace-Keeping Forces in Bosnia</EN-title>
<EN-desc>Reasons for the withdrawal of United Nations (UN) peace-keeping forces from Bosnia.</EN-desc>
<EN-narr>In 1994, some of the European nations participating in the Bosnian peace-keeping mission wanted to withdraw their forces. Relevant documents will report the reasons for the proposed withdrawal.</EN-narr>
<num>C049</num>
<EN-title>Fall in Japanese Car Exports</EN-title>
<EN-desc>Documents will report on the decrease in cars exported by Japan.</EN-desc>
<EN-narr>Documents reporting the fall in car exports from Japan in general or the decrease in car imports from Japan by a particular country or area are relevant. The decrease can be measured in terms of numbers of cars exported by or financial losses of the Japanese car industry. Documents that do not give some kind of figures to measure the fall in exports are not relevant.</EN-narr>

<num>C050</num>
<EN-title>Revolt in Chiapas</EN-title>
<EN-desc>Find reports on the uprising of Indians in Chiapas (Mexico).</EN-desc>
<EN-narr>Documents report on the reasons and the course of the rebellion of the indigenous population in Chiapas. They can also discuss the reactions of the Mexican government.</EN-narr>

<num>C051</num>
<EN-title>World Soccer Championship</EN-title>
<EN-desc>Find documents reporting on the final game of the world soccer championship of 1994.</EN-desc> <EN-narr>Relevant documents should give the results of the final game of the 1994 world soccer cup. Documents that discuss the final game in advance are not relevant.</EN-narr>

<num>C052</num>
<EN-title>Chinese Currency Devaluation</EN-title>
<EN-desc>Find documents describing the reasons and effects of the devaluation of Chinese currency.</EN-desc>
<EN-narr>Relevant documents discuss economic arguments in favour of and against the official reduction of the exchange value of the Chinese currency, and the social and economic consequences of the devaluation.</EN-narr>

<num>C053</num>
<EN-title>Genes and Diseases</EN-title>
<EN-desc>What genes have been identified that are the source of or contribute to the cause of diseases or developmental disorders in human beings?</EN-desc>
<EN-narr>A document that identifies a gene or reports that a gene has been discovered that is the source of any type of disease, syndrome, behavioral or developmental disorder in humans is relevant. Any document that reports the discovery of a defective gene that causes problems in humans is relevant, but reports of diseases and disorders that are caused by the absence of a gene are not relevant.</EN-narr>

<num>C054</num>
<EN-title>Final Four Results</EN-title>
<EN-desc>Find documents giving the results of the European Basketball Final Four.</EN-desc>
<EN-narr>Relevant documents will give details on the results of at least one of the three matches (two semi-finals and one final) of the final phase of the European basketball championship. Documents written prior to the semi-finals that give the names of possible winners are not relevant.</EN-narr>

<num>C055</num>
<EN-title>Swiss Initiative for the Alps</EN-title>
<EN-desc>Find documents that report on the Swiss initiative aimed at regulating traffic through the Alps.</EN-desc>
<EN-narr>New traffic rules aimed at respecting the ecological balance and encouraging green tourism are being introduced in Switzerland. Relevant documents will give information on the Swiss initiative in the Alps to regulate traffic in a sustainable way. </EN-narr>

<num>C056</num>
<EN-title>European Campaigns against Racism</EN-title>
<EN-desc>Find documents that talk about campaigns against racism in Europe.</EN-desc>
<EN-narr>Relevant documents describe informative or educational campaigns against racism (ethnic or religious, or against immigrants) in Europe. Documents should refer to organized campaigns rather than reporting mere opinions against racism.</EN-narr>

<num>C057</num>
<EN-title>Tainted-Blood Trial</EN-title>

<EN-desc>Find all information about the tainted blood trials in France including the sentences given by the court and the names of the people found guilty.</EN-desc>
<EN-narr>In 1994, there were several important trials in France prosecuting public health officials in connection with supplying contaminated blood to French hospitals. Relevant documents will give information on these trials and can also report details concerning the political consequences of this scandal.</EN-narr>

<num>C058</num>
<EN-title>Euthanasia</EN-title>
<EN-desc>Documents will describe incidents of euthanasia understood as "death with dignity" or "the right to die".</EN-desc>
<EN-narr>Documents describing any discussion relating to euthanasia interpreted as "death with dignity" or "the right to die" are relevant. Of interest are both reports of specific cases of euthanasia or discussions on legal aspects including current laws and moves to change the law. Documents which just mention "euthanasia" or the term "death with dignity" without giving details are not judged as relevant.</EN-narr>

<num>C059</num>
<EN-title>Computer Viruses</EN-title>
<EN-desc>Find documents about computer viruses.</EN-desc>
<EN-narr>Relevant documents should mention the name of the computer virus, and possibly the damage it does.</EN-narr>

<num>C060</num>
<EN-title>Corruption in French Politics</EN-title>
<EN-desc>Find documents on corruption in politics in France, in particular with reference to the illegal financing of French political parties.</EN-desc>
<EN-narr>Numerous important public figures in France, including politicians and leading industrialists, have been involved in episodes of corruption. Relevant documents will report such episodes. Instances of police investigations or court trials related to political corruption are also of interest.</EN-narr>

<num>C061</num>
<EN-title>Siberian Oil Catastrophe</EN-title>
<EN-desc>Find information on the rupture of an oil pipeline in Siberia.</EN-desc>
<EN-narr>Documents contain information on the rupture of an oil pipeline in Siberia, Russia, and its consequences. All technical and environmental aspects of this catastrophe are of interest, especially the long-term consequences for the ecological system of Siberia.</EN-narr>

<num>C062</num>
<EN-title>Northern Japan Earthquake</EN-title>
<EN-desc>Find documents that report on an earthquake on the east coast of Hokkaido, northern Japan, in 1994.</EN-desc>
<EN-narr> Documents describing an earthquake with a magnitude of 7.9 that shook Hokkaido and other northern Japanese regions in October 1994 are relevant. Also of interest are tidal wave

warnings issued for Pacific coastal areas of Hokkaido at the time of the earthquake. Documents reporting any other earthquakes in Japan are not relevant.</EN-narr>

<num>C063</num>

<EN-title>Whale Reserve</EN-title>

<EN-desc>Find documents about the reserve in the Antarctic in which hunting for whales is forbidden.</EN-desc>

<EN-narr>Relevant documents discuss the pros and cons of the Antartic whale sanctuary and mention countries that support the reserve or protest against it. Reports on violations of the protected area are also relevant.</EN-narr>

<num>C064</num>

<EN-title>Computer Mouse RSI</EN-title>

<EN-desc>Find documents that report on computer mouse repetitive strain injuries (RSI).</EN-desc>

<EN-narr>Relevant documents report injuries that are caused by the continuous use of a computer mouse. Documents proposing ways to avoid repetitive strain injuries (RSI) when using the computer are also relevant.</EN-narr>

<num>C065</num>

<EN-title>Treasure Hunting</EN-title>

<EN-desc>Find documents about treasure hunters and treasure hunting activities.</EN-desc>

<EN-narr>Identify types of current treasure hunting activities such as searching for gold, digging for buried relics, or searching underwater for sunken galleons.</EN-narr>

<num>C066</num>

<EN-title>Russian Withdrawal from Latvia</EN-title>

<EN-desc> Find reports and discussions about the withdrawal of Russian troops from Latvia.</EN-desc>

<EN-narr>Documents contain information on the discussion before, during and after the pullout of Russian troops from Latvia. They also include statements of Russian, Latvian, and other politicians on this action and the planning of this process. </EN-narr>

<num>C067</num>

<EN-title>Ship Collisions</EN-title>

<EN-desc>Find information on the number of people injured or killed in collisions between ships.</EN-desc>

<EN-narr>Relevant documents report information on the number of victims (dead or injured) of collisions between ships or naval vessels of all types. Documents that speak of victims without providing figures are not relevant.</EN-narr>

<num>C068</num>

<EN-title>Attacks on European Synagogues</EN-title>

<EN-desc>Find documents that describe acts of terrorism or vandalism against European synagogues since the end of the Second World War.</EN-desc>

<EN-narr> Relevant documents will mention bombings, attempted bombings or other acts of terrorism in or near synagogues in Europe. Also of importance are descriptions of profanities, threats or offensive writing on synagogue buildings. References to events previous to 1947 are not relevant.</EN-narr>

<num>C069</num>
<EN-title>Cloning and Ethics</EN-title>
<EN-desc>What are the practical applications of cloning, and what are the ethical arguments against it?</EN-desc>
<EN-narr>A document that reports on any practical application that cloning might have in everyday life is relevant, as are documents that report on moral/ethical arguments against cloning. Documents discussing generic bio-engineering or genetic engineering techniques are not relevant.</EN-narr>

<num>C070</num>
<EN-title>Death of Kim Il Sung</EN-title>
<EN-desc>Find documents giving biographical information on Kim Il Sung, the president of North Korea, who died in 1994.</EN-desc>
<EN-narr> Documents written after he passed away in 1994 that provide any kind of biographical information on Kim Il Sung or that give some history of his political activities are relevant. Documents written before his death reporting current events in which Kim Il Sung was involved are not relevant.</EN-narr>

<num>C071</num>
<EN-title>Vegetables, Fruit and Cancer</EN-title>
<EN-desc>Find documents that relate the eating of vegetables and fruit to cancer.</EN-desc>
<EN-narr> Documents reporting either positive or negative effects of eating fruit and vegetables on cancer are relevant.</EN-narr>

<num>C072</num>
<EN-title>G7 Summit in Naples</EN-title>
<EN-desc> What role was played by Russia in the G7 summit in Naples in 1994?</EN-desc>
<EN-narr>Relevant documents will mention the Russian objectives in participating and the role played by president Yeltsin in the G7 meeting in Naples in July 1994.</EN-narr>

<num>C073</num>
<EN-title>Norwegian Referendum on EU</EN-title>
<EN-desc>What were the reactions in the rest of Europe to the negative results of the Norwegian referendum in which Norway decided against membership in the European Union (EU).</EN-desc>
<EN-narr>Relevant documents report reactions to Norway's refusal to join the European Union, or discuss the consequences of this decision.</EN-narr>

<num>C074</num>
<EN-title>Inauguration of Channel Tunnel</EN-title>

<EN-desc>Find documents describing the inauguration of the Channel Tunnel and naming the national representatives of Britain and France present at this ceremony.</EN-desc>

<EN-narr>The Channel Tunnel project was financed by the French-British consortium TransManche Link. The opening ceremony took place on 6 May 1994. Relevant documents will report on the inauguration and give the names of the national representatives of Britain and France who were present.</EN-narr>

<num>C075</num>

<EN-title>Euskirchen Court Massacre</EN-title>

<EN-desc>Find documents on the court-house massacre in Euskirchen, Germany, in which 7 people died.</EN-desc>

<EN-narr>Documents report on the massacre in the small German town of Euskirchen where seven people were killed in the court-house. They can also speculate on the motivations of the killer.</EN-narr>

<num>C076</num>

<EN-title>Solar Energy</EN-title>

<EN-desc>In what applications is solar energy being used or being considered for future use?</EN-desc>

<EN-narr> Relevant documents are those that identify specific uses of solar energy systems. Documents that only describe the general technology of solar energy systems are not relevant.</EN-narr>

<num>C077</num>

<EN-title>Teenage Suicides</EN-title>

<EN-desc>What information is available concerning teenage suicides?</EN-desc>

<EN-narr>Only suicides of teenagers are relevant. Information regarding psychiatric care for depression in teenagers is irrelevant, unless a specific tie to teenage suicides is made. Teenage deaths which authorities suspect were accidental and not necessarily suicides are not relevant.</EN-narr>

<num>C078</num>

<EN-title>Venice Film Festival</EN-title>

<EN-desc>Which film or films won the Golden Lion in the 51st Venice Film Festival in September 1994?</EN-desc>

<EN-narr>Documents must contain at least the title(s) of the film(s) which was/were awarded the Golden Lion at the Film Festival in Venice. Golden Lions awarded for career achievements are not considered pertinent.</EN-narr>

<num>C079</num>

<EN-title>Ulysses Space Probe</EN-title>

<EN-desc>Find documents that describe the European space probe mission Ulysses or discuss its objectives.</EN-desc>

<EN-narr> Relevant documents will contain information on the Ulysses spacecraft and on the explorations for which it was employed, e.g. observations on the poles of the sun, studies on Jupiter.</EN-narr>

<num>C080</num>
<EN-title>Hunger Strikes</EN-title>
<EN-desc>Documents will report any information relating to a hunger strike attempted in order to attract attention to a cause.</EN-desc>
<EN-narr>Identify instances where a hunger strike has been initiated, including the reason for the strike, and the outcome if known.</EN-narr>

<num>C081</num>
<EN-title>French Airbus Hijacking</EN-title>
<EN-desc>Find all information concerning the role of an armed Islamic group in the hijacking of an Air France Airbus.</EN-desc>
<EN-narr>The Armed Islamic Group (GIA) have made numerous terrorist attacks in France. They were also responsible for the hijacking of an Air France Airbus. Relevant documents will report details on this hijacking.</EN-narr>

<num>C082</num>
<EN-title>IRA Attacks in Airports</EN-title>
<EN-desc> Find documents that describe terrorist acts by the Irish Republican Army (IRA) in European airports.</EN-desc>
<EN-narr>Relevant documents will mention shootings or other terrorist acts that the Irish Republican Army (IRA) has committed or threatened to commit in airports in Europe. Threats that have been revealed as false alarms are also relevant.</EN-narr>
<num>C083</num>
<EN-title>Auction of Lennon Memorabilia</EN-title>
<EN-desc>Find public auctions of John Lennon Memorabilia.</EN-desc>
<EN-narr>Relevant documents describe auctions that include objects that belonged to John Lennon or that are attributed to John Lennon.</EN-narr>

<num>C084</num>
<EN-title>Shark Attacks</EN-title>
<EN-desc>Documents will report any information relating to shark attacks on humans.</EN-desc>
<EN-narr>Identify instances where a human was attacked by a shark, including where the attack took place and the circumstances surrounding the attack. Only documents concerning specific attacks are relevant; unconfirmed shark attacks or suspected bites are not relevant.</EN-narr>

<num>C085</num>
<EN-title>Turquoise Program in Rwanda</EN-title>
<EN-desc>Find detailed information on operation ”Turquoise”, the French humanitarian program in Rwanda.</EN-desc>

<EN-narr>France initiated the operation "Turquoise" in south-west Rwanda during the conflict between the Hutus and Tutsis, in order to provide humanitarian aid to the population. Relevant documents will give information on this operation.</EN-narr>

num>86</num>
<EN-title>Renewable Power</EN-title>
<EN-desc>Find documents describing the use of or policies regarding "green" power, i.e., power generated from renewable energy sources.</EN-desc>
<EN-narr>Relevant documents discuss the use of renewable energy sources such as solar, wind, biomass, hydro, and geothermal sources. Low emission vehicles as for example electric or CNG cars are not relevant. Fuel cells are not relevant unless their fuel qualifies as renewable.</EN-narr>

<num>C087</num>
<EN-title>Inflation and Brazilian Elections</EN-title>
<EN-desc>Find documents analyzing the influence on the Brazilian elections of the "Plan Real" against inflation.</EN-desc>
<EN-narr>Relevant documents analyze the effects on the elections in Brazil of the "Plan Real" proposed by the Brazilian government to halt inflation.</EN-narr>

<num>C088</num>
<EN-title>Mad Cow in Europe</EN-title>
<EN-desc>Find documents that cite cases of Bovine Spongiform Encephalopathy (the mad cow disease) in Europe.</EN-desc>
<EN-narr>Relevant documents will report statistics and/or figures on cases of animals infected with Bovine Spongiform Encephalopathy (BSE), commonly known as the mad cow disease, in Europe. Documents that only discuss the possible transmission of the disease to humans are not considered relevant.</EN-narr>

<num>C089</num>
<EN-title>Schneider Bankruptcy</EN-title>
<EN-desc> Find documents on the bankruptcy of the German property speculator Schneider.</EN-desc>
<EN-narr> Documents report on the bankruptcy of the German property speculator Schneider and its background. They also examine the failures, omissions and responsibilities of the German banks in this case.</EN-narr>

<num>C090</num>
<EN-title>Vegetable Exporters</EN-title>
<EN-desc>What countries are exporters of fresh, dried or frozen vegetables?</EN-desc>
<EN-narr>Any report that identifies a country or territory that exports fresh, dried or frozen vegetables, or indicates the country of origin of imported vegetables is relevant. Reports regarding canned vegetables, vegetable juices or otherwise processed vegetables are not relevant.</EN-narr>