Cross-Lingual Document Retrieval, Categorisation and Navigation Based on Distributed Services

George Demetriou[a*], Inguna Skadiņa[d], Heikki Keskustalo[b], Jussi Karlgren[c], Daiga Deksne[d], Daniela Petrelli[e], Preben Hansen[c], Rob Gaizauskas[a] and Mark Sanderson[e]

[a]Department of Computer Science, University of Sheffield, UK
[b]Department of Information Studies, University of Tampere, Finland
[c]Swedish Institute of Computer Science, Stockholm, Sweden
[d]Tilde, Latvia
[e]Department of Information Studies, University of Sheffield, UK

[*]e-mail: G.Demetriou@dcs.shef.ac.uk

Abstract

The widespread use of the Internet across countries has increased the need for access to document collections that are often written in languages different from a user's native language. In this paper we describe Clarity, a Cross Language Information Retrieval (CLIR) system for English, Finnish, Swedish, Latvian and Lithuanian. Clarity is a fully-fledged retrieval system that supports the user during the whole process of query formulation, text retrieval and document browsing. We address four of the major aspects of Clarity: (i) the user-driven methodology that formed the basis for the iterative design cycle and framework in the project, (ii) the system architecture that was developed to support the interaction and coordination of Clarity's distributed services, (iii) the data resources and methods for query translation, and (iv) the support for Baltic languages. Clarity is an example of a distributed CLIR system built with minimal translation resources and, to our knowledge, the only such system that currently supports Baltic languages.

1. Introduction

The vast increase of multilingual content both on the Internet and corporate intranets has created the need for information access across languages and cultures. Cross Language Information Retrieval (CLIR) is a relatively new area of research and development that aims to overcome the cross-lingual access problem by enabling the users to retrieve documents written in one language - often called the *target* language - based on queries typed in another - often called the *source* or *query* language.

In this paper we discuss the architecture of the Clarity[1] system, a CLIR system that has been developed for English, Finnish, Swedish and Baltic languages. Clarity's main objectives were as follows:

- The development of a CLIR system that works with minimal translation resources, such as, for example, bilingual dictionaries.
- The implementation of retrieval methods that handle mixed collections of different language documents.
- The development of techniques of document organization and presentation that enable users to better interact with the system; such techniques include concept hierarchies and multi-document reports with filtering functionality.

A likely group of users of Clarity are *polyglots*, i.e. people who can speak or write in more than one language, who for one reason or another may feel comfortable in expressing and formulating queries in a single language, most probably their own native. For a really usable CLIR system, the requirements of the potential users need to be well understood. In section 2 of this paper we describe the user-centred design of Clarity that was based on observations and interviews with such a group of bi- & tri-lingual users working in the media, e.g. journalists and broadcasters. The user study led to the development of a user interface with novel functionalities for query input and disambiguation and influenced the way the system was built.

[1] Clarity's website is http://clarity.shef.ac.uk/

The requirement for CLIR systems to access and analyse data in different languages implies that they often need to integrate information from a variety of heterogeneous sources and software services at disparate sites. In section 3 we give an overview of Clarity's system architecture and we describe how different services are integrated into the system in order to perform a coordinated CLIR task.

Query translation is one of the fundamental elements of many CLIR systems, especially those that rely on a monolingual retrieval engine. The approach undertaken Clarity is query translation based on machine readable bilingual dictionaries. To deal with the problem of language pairs for which no direct translation exists we have adopted methods of transitive translation: the use of an intermediate language, usually termed a *pivot*, to provide the route between the source and target languages. The details of the implementation of the query translation algorithm as well the problems encountered in the context of the language pairs used in Clarity are summarized in section 4.

In section 5 we give an overview of the document organization and presentation facilities with particular reference to document clustering and filtering facilities that help the user browse, analyse and organise the retrieved document set.

The integration of Baltic languages posed specific challenges for the query translation and retrieval components of Clarity relating to the use of morphological information and the lack of direct translation routes from Latvian/Lithuanian to some of the other languages. In section 6 we describe the support mechanisms for the Baltic languages in Clarity and we provide retrieval results for the Latvian and Lithuanian for a range of system configurations (monolingual, cross-language direct, transitive and triangulated).

Finally, section 7 summarises what has been achieved and the lessons learned from Clarity's development.

2. User-Driven Interface Design

The requirements for both technical development and dialogue design were based on intensive user requirement analysis sessions, where interviews with professional information analysts and media professionals engaged in real-life multi-lingual tasks were combined with observational data from usage at workplaces. These were combined with our initial technical starting points to provide first versions of the interface including some of the planned functionality. These were tested using both laboratory experiments and renewed workplace visits in a user-centered iterative design cycle. The final dialogue design incorporates functionality as elicited from professional users, and also fulfils the goals of the research project -- to showcase the technological advances made during the course of the project.

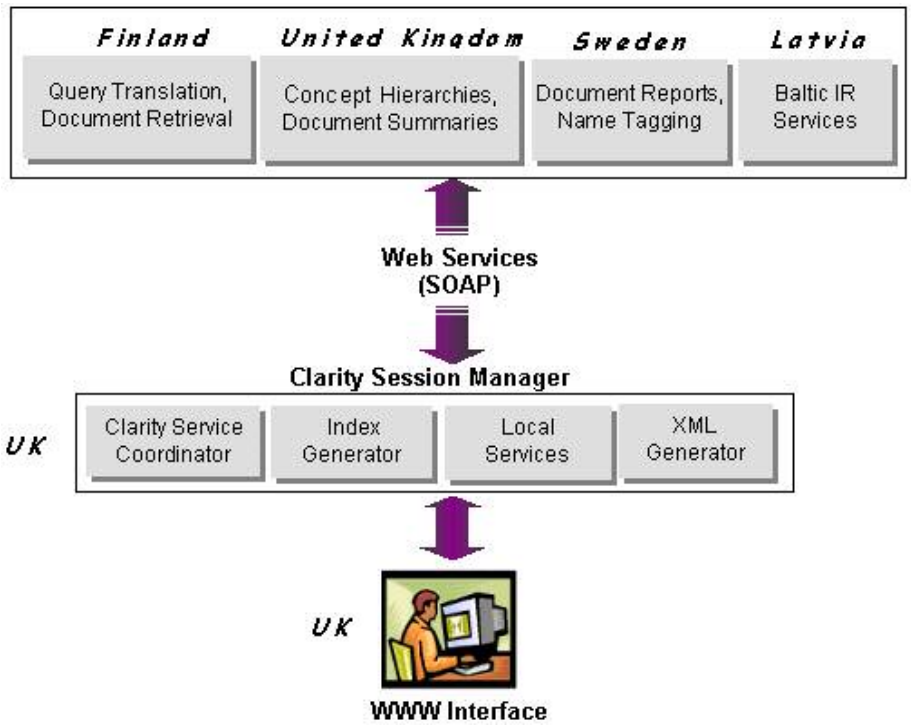

Figure 1: Clarity system architecture

3. System Architecture

Clarity has been designed in a way to provide flexibility for integrating new languages and textual databases and to promote the interoperability of system resources. To this end, Clarity's architecture has opted for a distributed model of services that allows for independent development of CLIR system modules, communication services and the user interface. The communication between system components is primarily facilitated by *Web services* (W3C 2002), i.e. services that are available on the Internet, use standardized messaging formats, such as XML, and enable communication between applications without being tied to a particular operating system or programming language. Clarity's three-layer architecture is shown in the diagram of Figure 1 with the *user interface* as a front-end, the data sources and services on the back-end and the *Clarity session manager* which is a middleware that 'sits' between the interface and the system services and provides the communication between them.

The back-end services support the following system functions:

- Query translation and multi-lingual document retrieval for English, Finnish and Swedish, Latvian and Lithuanian.
- Identification of query terms in retrieved document.
- Translation of target language document titles to source language.
- Extraction of the document's *best passage* for query terms.
- Document summarisation.
- Document clustering and presentation based on concept hierarchies (Sanderson and Croft 1999).
- Multi-document reports with filtering functionality that provide infromation in the form of text genre classification and named entity identification (Karlgren 1999).

Clarity's response times are satisfactory for interactive document retrieval. On the average, it takes 4 secs for the system to retrieve a batch of 10 documents for one target language. There is a relative increase in the response times when more target languages are searched but the retrieval times are still within acceptable bounds (7 secs for two target languages, 11 secs for three target languages). The impact of multi-user usage was found to be moderate and Clarity can support interaction for up to five users simultaneously without considerable overhead in response times (11 secs for 5 concurrent clients).

4. Query Translation

Multilingual collections can be accessed by letting the user express the search topic in another language than the target collection. For example, by using a Latvian query, the user may wish to retrieve English documents, or vice versa. The UTACLIR framework is used in Clarity for the query translation (see Hedlund et al., 2004). The framework utilizes external linguistic resources and query key is processed according to its perceived type.

External resources used are the following. Morphological analysis is utilized for transforming the source and/or the target keys into basic forms, and for splitting the source compounds into components if necessary (Lingsoft's *TWOL* for Finnish, Swedish, and English). Machine readable dictionaries are used for translating the source keys into the target language (Tilde's Baltic dictionaries for Latvian/Lithuanian and English - in both directions; via SOAP services; and Kielikone Inc.'s GlobalDix for 18 language pairs). Approximate string matching methods are used for finding the most similar words from the target database index, in case of untranslatable keys. Finally, stop word lists are needed for removing selected source and target language words during the translation process.

The following source key types are recognized and processed specifically by the current translation framework:

(1) Stop words: non-content bearing words belonging to used-defined lists will be omitted. Translating source stop words - or keeping target stop words as part of translations - would add unnecessary noise into the target query.

(2) Recognized translatable words: source keys belonging to this category are such which are both recognizable (included in the lexicon of the morphological analyzer) and translatable (included in the translation

dictionary). These words will be translated, and the translations will be treated as synonyms in the target language (connected with synonym operator). Typically, these words include all the most common words of the language

(3) Recognized untranslatable and unsplittable words: source keys belonging to this category can be recognized by the morphological analyzer, but they are untranslatable, and they cannot be splitted by the morphological analyzer. Typically, this kind of words include proper names and rare terminology, and they occur because the relatively large lexicon of the morphological analyzer enables the lemmatization of words yet missing from the translation dictionary.

(4) Recognized untranslatable but splittable words: source keys belonging to this type include compounds not included in the translation dictionary as whole words. This is a vast class of words because any number of novel compounds may be created "on the fly" in compound languages (e.g., Finnish or Swedish). Thus, this constitutes an important key case.

(5) Unrecognized but translatable words: source keys belonging to this type are rare, because the translatable words are typically also recognizable by our morphological analyzers used. However, if such keys exist, they are translated.

(6) Unrecognized and untranslatable words: source keys belonging to this type are unrecognized by the morphological analyzer and also not translatable. This kind of words include typically proper names, acronyms, scientific terms, rare words and new words of the language. As direct translation is not possible, approximate matching is performed to find the most similar strings occurring in the target index.

(7) Numbers: keys expressed by digit strings are used as such also in the target language.

(8) Keep as-is words: whenever the user precedes the query with a special symbol, the translation is not performed, but the user given key is simply copied as such into the target query.

(9) Enforced fuzzy keys: by using a special symbol (~) the user defines a "fuzzy key". This means that the most similar words are retrieved from the target index and placed into the target query regardless of whether the key could be translated or not. There are several interesting applications for the fuzzy key. For example, "problematic" words, such as proper names, misspellings, and technical terms not occurring in the translation dictionary could be handled this way.

At the implementation level, the query translation program utilizes a simple tree data structure. The uppermost level of the tree consists of the original source keys given by the user, and it also reflects the logical structure of the original source query. The second level nodes contain the processed source language strings, for example, normalized forms generated by the morphological analyzer. The final third level of the tree consists of post-processed word translations (in the target language). Once built, the tree structure can be traversed and interpreted in different ways, for example, structured or a list type of translation can be selected.

The basic query translation system performs direct query translation. However, transitive translation can also be performed by translating from source language to target via a pivot language (Gollins & Sanderson, 2001). In our implementation, the pivot result is formed as a simple word list while the target query is formed as structured, that is, containing synonym operators. Moreover, a triangulated translation can be performed by using two different transitive routes and by combining the result into one. For Baltic languages, Clarity supports direct query translation routes English-to-Latvian, English-to-Lithuanian; and transitive translation Finnish-English-Latvian and Latvian-English-Lithuanian; plus triangulated translation from Finnish to Latvian via two separate pivot languages, English and German.

As an example, in triangulated translation from Finnish to Latvian, the Finnish source query 'vakoiluskandaalissa ames' is translated into the following target query:

```
#sum(#or(#syn(spiegošana skandāls
negods skandalozs apkaunojšs)
#syn(spiegošana @belle blefs
negods @schreder @gathered
skandāls lērums mēlnesīgs))
#or(#syn( pames samest @aames
aames)#syn(pamest samest @ames
ames)))
```

The example above illustrates the complexity of the triangulated translation task. The first source key, an inflected form of a Finnish compound *vakoiluskandaali* (expressing concept *espionage scandal*) – is processed first by Fin-Eng and Fin-Ger direct translation steps. The Finnish key belongs to a key type "*recognized*" (by morphological analysis). Moreover, it is classified as being an *untranslatable* word (because it is not found from the translation dictionary as a whole) and also a *splittable* (through Finnish morphological analysis). Thus, it is splitted and its normalized components are translated (by default) by the triangular system into both English and German. Secondly, the results of both these translations are next translated into Latvian by using Eng-Lat and Ger-Lat (respectively) direct translation steps. In the output, as an example, the very first synonym (#syn) set inside the first #or statement is derived from the English-Latvian translation corresponding to the original Finnish compound *vakoiluskandaalissa.* The second synonym set immediately after it is derived from the German-Latvian translation (corresponding to the same original Finnish compound). The second Finnish source key *ames* is untranslatable. Fuzzy matching translation features are used as part of the component processes.

In the example above (as in the present prototype), we have decided to combine the corresponding translations with the InQuery's #or operator, but other operators (#syn etc.), and more generally, other query structuring options can be used.

5. Document Organisation and Presentation

A common presentation format in information retrieval systems is a list of documents (or document titles) sorted by some sort of relevance ranking as understood by the system. In Clarity the retrieved list is the default first view of retrieved items but, as a complement, the results can be presented in the form of a concept hierarchy, a textual retrieval report or a list of document summaries.

In concept hierarchies, documents are clustered with respect to a hierarchy of concepts that are derived from the set of retrieved texts; the resulted structure is presented as a set of hierarchical menus according to the statistical principle of 'subsumption' (Sanderson and Croft 1999). The appealing characteristic of 'subsumption hierarchies' is that they can be automatically extracted without the need for prior knowledge or training data. In Clarity target language documents are organized into clusters of source language concepts. An example of a concept hierarchy in English generated from Latvian retrieved documents is shown in Figure 2.

The second organisation facility is a multi-document report that describes the retrieved set in terms of various extracted features of individual items: number or frequency of search terms (see previous sections); alternative keywords, names and other data extracted from the documents; language; and text genre (e.g. news article, opinion piece, interview) or style of text e.g. argumentative, subjective, personal etc. (Karlgren 1999). These informational elements can be activated as filters to help users reduce the retrieved set to manageable proportions before inspecting individual documents.

A third option provides users with summaries of the original document. This has been developed as a facility to assist users get a view of a document's content quickly without the need to read the whole document.

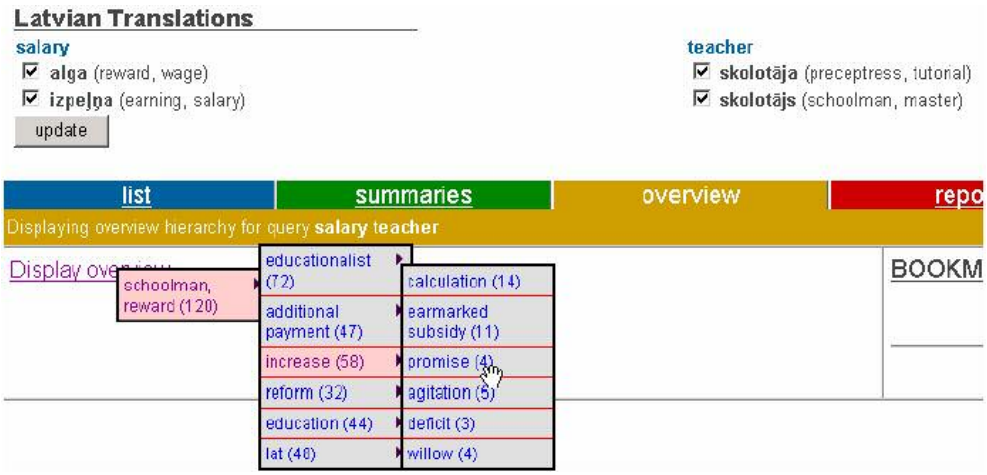

Figure 2: English hierarchy generated from Latvian documents for the query 'teacher salaries'.

6. Baltic languages in cross-language information retrieval systems

One of the aims of the Clarity project was to develop methods that allow adaptation of CLIR techniques to new languages that have no expensive translation resources, so called low-density languages.

In 2002, when Tilde joined Clarity project, monolingual and cross-lingual information retrieval (IR) was relatively rare field of studies for Baltic languages. In this time systems which treated inflected forms of Baltic languages were not developed. Therefore at first morphological analysis constituents necessary for monolingual retrieval was adapted and included into indexing engine of the Clarity system, allowing retrieval of inflected forms by single normalized key.

Although monolingual retrieval was not between the main objectives of the project, automatic tests on test collections and user tests were performed for monolingual retrieval also. The tests showed that the best results are obtained in cases when the query is formulated clearly. For instance, in Latvian best results showed queries *eitanāzija (euthanasia), genoma loma medicīnā (genome role in medicine)* and *datorvīrusi (computer viruses)*, while for topic *kapitālieguldījumi Austrumeiropā vai Krievijā* (investments in Eastern Europe or Russia) average precision was zero.

Inclusion of Baltic languages into the cross-language information retrieval system is realized through the UTACLIR framework. A translation SOAP service for following language pairs was developed: English-Latvian, Latvian-English, German-Latvian, Latvian-German, Russian-Latvian, Latvian-Russian, Lithuanian-English, English-Lithuanian, German-Lithuanian, Lithuanian-German, Lithuanian-Russian and Russian-German. Baltic language translation services are used for two purposes: query translation and headline translation.

Query translation in Clarity system is performed in two directions: from Baltic languages into English, Finnish and Swedish to retrieve documents in these languages; from English, Finnish and Swedish into Baltic languages for retrieval in Baltic language document collections. Since there are not direct translation dictionaries between Finnish/Swedish and Latvian/Lithuanian, transitive query translation was performed in these cases. Two transitive retrieval methods have been research: simple transitive retrieval through English and triangular transitive retrieval where English and German are used as pivots.

One of most complicated issues was selection of translations during query translation process: Tilde's electronic dictionaries contain rich content; most of words have several translations or even several meanings. Although UTACLIR has mechanism allowing rich usage of synonyms to avoid overgeneration only the first two meanings are used for query translation and cross-language retrieval.

To estimate the quality of CLIR for Baltic collections, tests with 45 topics were performed for monolingual and cross-lingual retrieval. First results are very promising, especially when it concerns transitive retrieval: average precision for English-Latvian retrieval is 8.8 % (or 46.32% from monolingual retrieval) and 26.3 % (or 72.85% from monolingual retrieval) for English-Lithuanian retrieval. Surprisingly high average precision 28% is obtained for transitive retrieval between Latvian and Lithuanian using English as pivot language. Summarised results for the Latvian test collection (with the 45 topics expressed in Latvian, English and Finnish relevance judgments) are given in Table 1.

Target: Latvian	**Average Precision (%)**
Monolingual baseline (Latvian queries)	19.0
Direct query translation (Eng->Lat)	8.8
Transitive query translation (Fin->Eng->Lat)	8.1
Transitive query translation (Fin->Ger->Lat)	8.7
Triangular query translation (with #or(#syn(…)) sets)	9.1

Table 1: Cross language retrieval results for Latvian language.

To generate these results we have compared monolingual retrieval with direct translation, two transitive translations and one triangulated translation in a laboratory model setting. As query words, we selected the title field words of the topics only (short "brief search" type queries, typically around 3 query words). As can be seen from Table 1, the translation methods reached around 40 % of the monolingual results; however, further studies may be needed to explain and improve the Baltic translation results.

Since queries in test collections were not always precise, which is illustrated by low precision of monolingual retrieval, user tests of Latvian-English-Lithuanian transitive retrieval were also performed. The user was asked to test topics which in automatic tests had very low or very high precision. In general, the user was satisfied with system and pointed main reasons why for some topics retrieval showed poor results. Reasons mentioned by the user are: synonyms used in text; generality of topic (if the topic is *floods in Europe*, user needs to ask for *floods in Germany, floods in Poland,* etc.) and in some cases quality of transitive retrieval was affected by retrieval errors (word *kapitālieguldījumi-investments* lost its main meaning during transitive translation process).

With respect to document presentation, the technique of concept hierarchies was adapted for Baltic languages and used for monolingual and cross-lingual retrieval. When the query language is one of Baltic languages the concept hierarchies are built in corresponding Baltic language. If the query language is English then the hierarchies are built in English through several steps. At first query is translated into Latvian or Lithuanian and requested documents are retrieved. As a next step, Latvian terms are extracted from the top 200 retrieved documents and translated into English. Term selection is based on following principles: 1) we take the nouns from the closest two sentences where any form of any query word is found; 2) if there are two following nouns and the first of them is in genitive form we take these two nouns as a noun phrase, for example, "*bank's president*"; 3) if there are two following nouns and the second of them starts with capital letter we take the nominative forms of these two nouns as a noun phrase, for example, "*teacher John*"; 4) from selected term list we remove the terms which statistically are too common. Afterwards, term subsumption is calculated.

7. Summary and Conclusions

In this paper we have discussed the Clarity system, a CLIR system for English, Finnish, Swedish and Baltic languages.

We have built Clarity on a platform based on Web services and our experience is that it helped the system developers to deal with data licensing issues, to avoid re-installing software implementations and to concentrate on the localization of resources, thus reducing overall development time. Performance-wise, the results indicate that Clarity can support CLIR tasks in a timely manner for a reasonable number of users simultaneously. We have also learned, however, that computationally intensive tasks, such as the generation of concept hierarchies or document reports are better supported by services at a single site where all the associated data resources are readily accessible possibly in a precompiled form. For example, because the available processing resources for the extraction of named entities from texts could not guarantee the real-time generation of document reports, the decision was to annotate the document collections in advance and store the results in database indexes for fast access. We have also found that the Web services approach was a good choice for overcoming problems relating to character set compatibility between the Baltic and the other languages such as between Latin-7 and Unicode.

With respect to query translation, we have applied methods that proved successful in dealing with the incompleteness and ambiguity of bilingual dictionaries. Dictionary incompleteness was dealt with by mechanisms that exploit the constituency structure of compound terms and a fuzzy match algorithm which is used to propose translation candidates for untranslatable terms. Although such mechanisms may result in an increase in the number of possible translations for a query, we have found that the problem is greatly minimized by allowing the users to select the translations on the interface they think are most relevant. For operations in which the translation is done implicitly by the system, such as the translation of terms in titles or

concept hierarchies, the selection of the two or three top ranked translations proved to be an effective method for eliminating the 'noise' due to irrelevant translations. However, further studies are needed to reveal the extent to which such a pruning may result in the removal of also relevant translations.

With respect to Baltic languages, the results for document retrieval using direct query translation indicate that the average precision can reach levels of more than 70% compared to monolingual retrieval. In the case of transitive translation the precision is lower (which is to be expected due to the extra noise from translations introduced by the pivot languages) but still at reasonable levels compared to monolingual (around 40%). Preliminary tests with triangulated translation showed a small improvement over transitive translation but further studies may be needed to reveal whether this is significant.

Overall, we believe that the development of Clarity can serve as a methodology for building practical, usable CLIR systems with current technologies and limited data resources.

8. Acknowledgements

The following resources were utilized by UTACLIR. ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright © 1989-1992 Atro Voutilainen and Juha Heikkilä. FINTWOL (Morphological Description of Finnish): Copyright © Kimmo Koskenniemi and Lingsoft Oy. 1983-1993. GERTWOL (Morphological Transducer Lexicon Description of German): Copyright © 1997 Kimmo Koskenniemi and Lingsoft, Inc. GlobalDix and MOT Dictionary software. Copyright © 1998-2002 Kielikone Oy, Finland. SWETWOL (Morphological Transducer Lexicon Description of Swedish): Copyright © 1998 Fred Karlsson and Lingsoft, Inc. TWOL-R (Run-Time Two-Level Program): Copyright © Kimmo Koskenniemi and Lingsoft Oy. 1983-1992.

9. References

Gollins T., Sanderson M. (2001) Improving Cross Language Information Retrieval with Triangulated Translation. SIGIR 2001: 90-95.

Hedlund T., Airio E., Keskustalo H., Lehtokangas R., Pirkola A., Järvelin K. (2004) Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. Inf. Retr. 7(1-2): 99-119.

Karlgren J. (1999), Stylistic Experiments in Information Retrieval. In: Natural Language Information Retrieval. Tomek Strzalkowski, (ed.), Kluwer.

Sanderson, M. and Croft, W.B. (1999). Deriving concept hierarchies from text. In Proceedings of the 22nd ACM SIGIR Conference, 206-213.

World Wide Web Consortium – W3C (2002a). Web Services Activity. At: http://www.w3.org/2002/ws/.