UNIVERSITY OF TAMPERE

This document has been downloaded from
Tampub – The Institutional Repository of University of Tampere

Post-print

Tekijät: Airio Eija
Artikkelin nimi: Who benefits from CLIR in web retrieval?
Julkaisuvuosi: 2008
Lehden nimi: Journal of Documentation
Lehden vol: 64
Lehden numero: 5
Sivunumerot: 760-778
Tieteenala: Luonnontieteet / Tietojenkäsittelyoppi ja informaatiotieteet
Kieli: en
URN: http://urn.fi/urn:nbn:uta-3-472

All material supplied via TamPub is protected by copyright and other intellectual property rights, and duplication or sale of all part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorized user.

Who benefits from CLIR in Web retrieval?

Keywords Cross-language information retrieval, User tests, Web

Research paper

Purpose
The aim of the current paper is to test whether query translation is beneficial in Web retrieval.

Design/methodology/approach
The language pairs were Finnish-Swedish, English-German and Finnish-French. We recruited 12-18 participants for each language pair. Each participant performed four retrieval tasks. Our aim was to compare the performance of the translated queries with that of the target language queries. Thus, we asked participants to formulate a source language query and a target language query for each task. The source language queries were translated into the target language utilizing a dictionary based system. In English-German, also machine translation was utilized. We used Google as the search engine.

Findings
The results differed depending on the language pair. We concluded that the dictionary coverage had an effect on the results. On average, the results of query-translation were better than in the traditional laboratory tests.

Value
This research shows that query translation in Web is beneficial especially for users with moderate and non-active language skills. This is valuable information for developers of CLIR systems.

1 Introduction

The traditional laboratory IR model is based on a test collection and test topics with their binary relevance assessments. Test queries are usually formulated automatically from test topics. The performance of a system or a query is measured by recall and precision. The model is very controlled, enabling easy comparisons across various test situations. However, retrieval systems are designated for human users. Thus, user-oriented research is needed besides laboratory tests.

The sizes of the traditional research collections are quite small. Hull asked in 1993 whether experimental results with small collections and short documents are applicable to large collections. He answered that there is no evidence that they would not be, but researchers must be cautious when generalizing specific test result into different environments. (Hull 1993.) The test collection sizes have increased since that, but so have the real world collection sizes. There has been the terabyte track in TREC since

2003. The collection, GOV2, contains about 25 million documents (426 GB) from .gov pages (see http://www-nlpir.nist.gov/projects/terabyte/). In 2005, Google claimed to index over 8 billion pages and MSN about 5 billion pages. Gulli and Signorini estimated that the number of the indexable Web pages was at least 11.5 billion in the beginning of the year 2005. (See Gulli & Signorini 2005.) The number of the GOV2 documents is about an eight hundredth part of the number of the Google documents. In addition to the size, diversity characterises the Web. We can ask whether the GOV2 collection really stands for the whole Web: all documents genres as well as languages. English dominates the GOV2 collection, and thus it is not the best possible test collection for non-English retrieval.

Most of the interactive cross-language information retrieval (CLIR) experiments have been performed with test collections (see e.g. iCLEF - interactive track for the Cross-Language Evaluation Forum; Petrelli et al. 2006). There is not much research on the applicability of query translation with large collections, or in connection with existing Web engines. The performance of translated queries compared with target language queries in the Internet has not been tested widely, either.

Ogden with colleagues (1999 and 2000), Capstick with colleagues (2000) and Penas with colleagues (2001) have reported of development of Web-based CLIR systems. Their research has concentrated on operation and interaction of the systems, and they did not perform any user tests concerning the performance of the systems.

This study focuses on possible benefits of query translation in the Web for real users. Query translation (without document translation) might be useful for searchers with good or moderate target language skills: those with poor skills are not able to read the documents. We assume that users with moderate / passive target language skills benefit most of query translation: they are able to read documents, but they might have difficulties in formulating target language queries because of problems with the lexicon and / or the orthography. Query formulation is presumably easier for those with good / active skills, and thus they would not benefit from query translation so much. We have recruited participants with moderate and good target language skills for our user test in order to test whether our assumption is right. There is not much prior research on the usefulness of query translation for users with different language skills, although it is important to identify the potential target group when real CLIR systems are developed.

The structure of this paper is the following: Section 2 discusses research on interactive CLIR and CLIR systems in the Internet. The settings of this test as well as

research questions are presented in Section 3. Section 4 describes the results and Section 6 contains the discussion and conclusions.

2 Previous research

2.1 CLIR user tests

Laboratory tests dominate IR and CLIR research, but there are attempts toward interactivity. There has been an interactive track in CLEF (Cross Language Evaluation Forum), iCLEF, since 2001. CLIR user tests were performed also in connection with the CLARITY project (Cross Language Information Retrieval and Organisation of Text and Audio Documents) (Petrelli & al. 2006).

In the first iCLEF, the research task was to compare two CLIR systems regarding their ability to inform the users about the document contents. There were two possible document collections, English and French.

In iCLEF 2002, the research task was to compare two systems differing in at least one of the following aspects: support for document selection, support for query translation or support for query refinement. There were three possible document languages, while topics were available in twelve languages. The purpose was to use the topics in the native language of the searchers. A standard example of instructions was given in order to standardize the experiment conditions. Information was gathered also by standardized questionnaires. The searchers should answer questions at the start of the session, after each topic, when switching systems, and at the end of the session.

Participants of iCLEF 2003 were supposed to concentrate on some of the following aspects: formulating and / or translating the query, refining the formulation and / or translation of the query and identifying foreign-language documents as relevant. The task was to compare two systems, which differ in some of the aspects listed above. If the focus was on query formulation, translation or refinement, the search task consisted in finding as many relevant documents as possible. If the focus was on document selection, the search task consisted of scanning a fixed ranked list of documents and selecting the relevant ones.

In 2004, iCLEF focused on the problem of Cross-Language Question Answering (CL-QA) from a user-inclusive perspective. The task was twofold: 1) how well did the system help users to locate and identify answers to a question, and 2) how well did the interaction with the users help a cross-language QA system to retrieve better answers.

The users should belong to either of these groups: searchers with passive language abilities in the foreign language, or searchers with no useful language abilities in the foreign language. The settings in iCLEF 2005 were approximately similar with iCLEF 2004. (iCLEF interactive track for the Cross-Language Evaluation Forum)

Petrelli and colleagues tested in 2006 two kinds of CLIR user interfaces with the Clarity project: a fully delegated interface, where the user types in a query and gets a document list; and a supervised interface, where the user is presented with a list of translations for each query term along with their appropriate back-translations in parentheses. The translations were arranged in columns. The user could de-select translation alternatives and insert new ones before clicking the *search* button. (Petrelli & al. 2006.)

The participants were either students or academic professionals. In the beginning of the tests, participants received a written briefing on the purpose and procedure of the test. They also filled in a questionnaire about personal information (education, age, etc.). Each participant used both of the two interfaces under testing. Users conducted two similar tasks with different interfaces. After each task, they filled in questionnaires concerning their familiarity with the topic, as well as user satisfaction. After the tasks had been completed, the users filled a questionnaire addressing systems comparison: which one was easier to learn and use. (Petrelli & al. 2006.)

The authors found that users were not interested in, or were not able to, control the query translation steps. Graphical visualisations of the result did not interest them either. Users wanted an interface where they could type in a query and receive a list of relevant documents. (Petrelli & al. 2006.)

Thus, the purpose of both iCLEF and CLARITY user tests has been to test two CLIR systems differing in some aspect. They have not tested the performance of a CLIR system compared with a monolingual system. Both iCLEF and CLARITY user tests have been performed in test collections.

2.2 CLIR web search engines

There is also research on developing a functional cross-language web search engine. Capstick and colleagues presented their cross-language web search engine MULINEX, and Bian and Chen their query and document translation system (MTIR) in 2000 (see Bian & Chen 2000; Capstick & al., 2000).

In MULINEX, the user types in a query and selects the source language and the target languages from a menu. The query is translated into the target languages, and the translations with back-translations are shown to the user. The user selects desired translation alternatives, and presses the *search* button. The result list contains documents in selected languages sorted by their estimated relevance. For each document, the language, title, URL, size, category and summary are displayed. The summary is presented in the document language, but it can be translated. (Capstick & al., 2000, 4-6.)

MULINEX is based on a database containing information about all the documents. A document may be composed of several web pages in a frameset. Queries were translated using bilingual dictionaries. Summaries were translated utilizing a machine translation system. (Capstick & al., 2000, 6-12.)

MTIR utilizes both dictionary based and corpus based approaches in query translation. For proper name searching, they developed a machine transliteration algorithm. For document translation, several applications were developed, for example an analysis module for analyzing the structure of the sentences and a synthesis module for dealing with word insertion, deletion and refinement. (Bian & Chen 2000.)

MULINEX and MTIR differ in many aspects. MTIR handles Chinese-English translation, while MULINEX has more language alternatives. MTIR was developed both for query and document translation. MULINEX translates queries. Summaries are presented in the document language, but they can be translated. (Capstick & al., 2000)

3 Test settings and research questions

The previous section showed that in prior CLIR research, user tests have been performed mainly in test datasets. Systems for query translation in the Web have been created, but their usefulness has not been tested widely. Thus, there is not much research on the applicability of CLIR in the Web.

The main purpose of the current paper is to study whether query translation is advantageous in Web information retrieval for users who are able to read documents written in the target language. We do not handle document selection, because our focus group is able to read and understand retrieved documents. Our focus group consists of users with moderate language skills: they are able to read target language documents, but query modification might be difficult for them. For comparison, we also had active

target language users among our test persons. Presumably, they do not benefit from query translation very much. Instead, we did not recruit participants with poor target language skills: they might not be capable of reading documents. The participants are detailed in Section 3.3.

We focused on three language pairs: Finnish-Swedish, English-German, and Finnish-French. We had 18 Finnish-Swedish, 12 English-German and 12 Finnish-German participants in our tests. We wanted to simulate a real world retrieval situation. The most authentic option would be to utilize topics created by each participant. On the other hand, result comparison would not be possible in that case. Giving the same topics for each participant would not have been realistic either, because users have differing interests. We concluded to introduce ten search tasks, out of which we asked participants to select four. We asked them to formulate their queries both in the source language (Finnish or English) and in the target language (French, Swedish or German). We translated automatically the source language queries into the target language, and compared the performance of the user formulated target language query with the performance of the translated query to test which one performed better.

3.1 Research questions

Our research questions are the following:

1. What is the relative performance of the users' direct querying in the target language vs. the automatically translated target language queries? Which performs better, dictionary based translation or the user formulated target language query
    a. in Finnish-Swedish user tests?
    b. in Finnish-French user tests?
    c. or machine translation, in English-German user tests?
2. Does the performance of the translated queries / the target language queries differ significantly in the groups according to
    a. the language skills?
    b. the topic domain familiarity?
    c. the topic vocabulary familiarity?

3.2 Test settings

We compiled an interface where the user should formulate two queries: one in the source language and one in the target language. The source language query was automatically translated into the target language. By comparing the performance of these two queries – the user formulated target language query and the translated query – we could make conclusions concerning the usefulness of CLIR for the user. The test settings differed from the normal retrieval circumstances: users do not usually formulate two queries per search. The settings were admittedly artificial, but a step towards the real world and the real users compared with the traditional laboratory CLIR tests, where queries are formulated automatically out of topics, standard relevance corpora (often binary) are utilized, and tests are performed in restricted test datasets. In the current test settings, the users formulated their queries themselves, performed retrieval in a real environment (the Web), and evaluated the documents themselves using the multi-graded relevance scale. Thus, this kind of test gives more truthful information about real users and their actions in a realistic environment than the traditional laboratory tests.

The circumstances in a real life IR may vary: the user might have (traditional or electronic) dictionaries in use or not, he / she might use the possible dictionaries or not, and the user might be ready to perform multiple searchers for one topic or not. We did not allow participants to utilize any dictionaries in the test: the aim of CLIR is to help users in query formulation so that no dictionaries are needed, and we wanted to simulate that situation. In addition, those fluent in the target language would hardly use any dictionaries at all. We allowed the participants to perform only one search per one task. If we had allowed more, the participants might have learned vocabulary from the documents retrieved with the translated query. That could have helped the participants to formulate target queries, which would have distorted the test setting. The search topics are described in Section 3.4.

The retrieved lists of both query versions were merged prior to presentation to the user. The system and the interface are described in Section 3.5.

The participants were asked to assess the relevance of the retrieved documents in four-graded relevance scale.

3.3 Participants

We recruited participants among students at the University of Tampere: for Finnish-Swedish and Finnish-French tests students whose mother tongue is Finnish, and for

English-German students who are fluent in English. We selected participants who have taken a maturity examination in the target language and are able to read texts written in the language. We did not recruit students of information retrieval, because their advanced abilities in query formulation might have affected the results. We asked the participants to fill in a short questionnaire about their background (see Appendix 1).

The Finnish high-school graduation grades are a, b, c, m, e and l in the ascending order. We converted them to numbers 5-10. The high school report grades vary between 5 and 10 as well. In order to evaluate the influence of the language grades on the results of a single participant, we summed the report grades and graduate grades, and classified the sums into *two language grade* groups: 1 point (scores 10-17) and 2 points - (scores 18-20).

The frequency of using the target language affects language skills besides the grades. We interpreted here reading and listening as passive language practice, while writing and speaking were interpreted as active practice. We asked the participants how often they had read or listened the target language during the last two years, and how often the they had spoken or written it (1 - not at all, 2 - a little, 3 - monthly, 4 - much). Target language query construction requires active language skills for good results. We multiplied the active language grade by two, summing the product with the passive language grade, and classifying the sums into two *language use* groups: 1 point (scores 3-6) and 2 points – (scores 7-12).

We calculated the *language skills measure* by summing the *language grade* points with the *language use* points and classifying the sums into two groups: 1- moderate skills (2-3 points) and good skills (4 points).

We recruited 12-18 participants for a language pair and asked each of them to select four search tasks. Thus, we have at least 48 results for a language pair. Participants had 30 minutes for each task. They received a free movie ticket for their participation in the test.

3.4 Tasks

In the simulated work task approach, the participants are given a short, quite open cover story describing a fictive situation leading to retrieval (see Borlund 2000). We presented our participants a short cover story for each topic which motivated them for retrieving documents in the target language. We had ten topics for each language pair (see

Appendix 4 for English-German topics), all concerning quite general matters. Participants were told to select four most interesting among the topics. After performing each task, participants were asked to fill in a questionnaire about their familiarity with the topic subject area and the target language vocabulary concerning the topic (see Appendix 2). Thus, we had two measures for the tasks: *a topic familiarity measure* and *a vocabulary familiarity measure*, which both had three values: 1 - not at all familiar, 2 - a little familiar, 3 - very familiar. After performing all the tasks, we asked the participants to answer question concerning the queries and the assessment process (see Appendix 3).

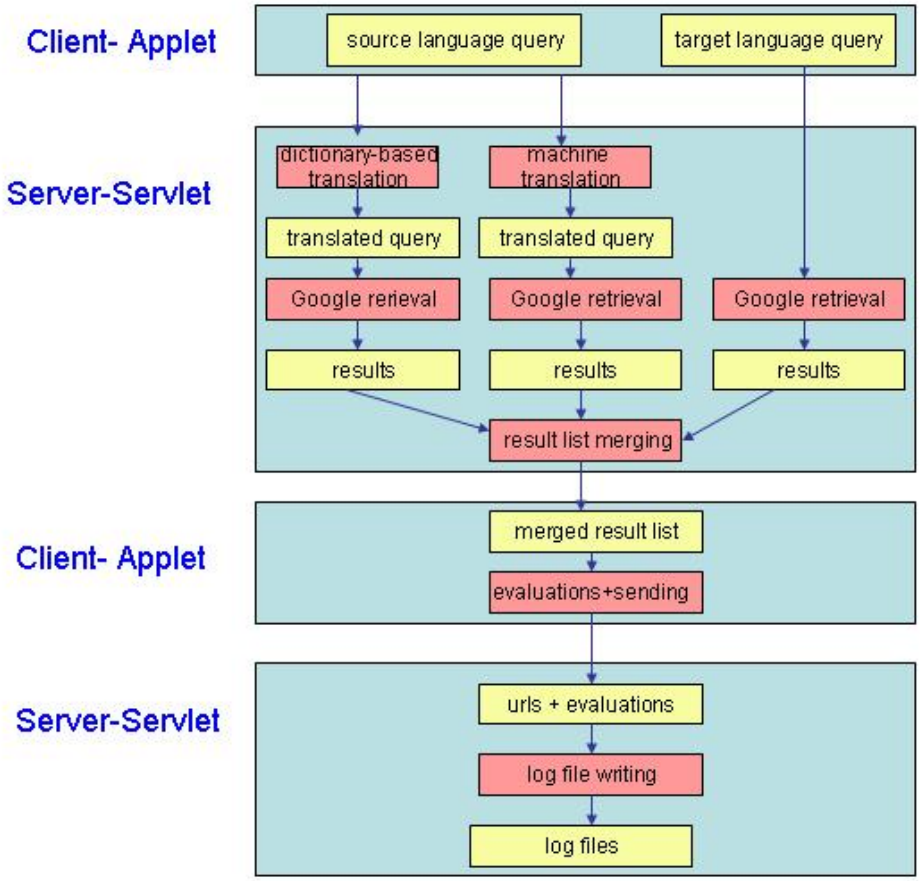

Fig. 1. Process overview of the query translation and evaluation system

3.5 Systems and queries

We used two kinds of query translation systems in the experiments: a dictionary based translation system UTACLIR (developed at University of Tampere), and the publicly available MT system Babelfish. These were incorporated under our experimental CLIR

interface which also supported direct target language searching. UTACLIR was utilized with all the language pairs and Babelfish with English-German translation.

The UTACLIR framework supports the use of external language resources. First, source language word is normalized utilizing a lemmatizer. Stop words are removed after lemmatization, and non-stop words are translated. Synonyms as well as parts of a compound are enveloped by the synonym operator (which was here converted to an OR operator according the Google syntax). Translated words are lemmatized or stemmed, depending on the target index. Here, we used lemmatization. Last, target stop words are removed. Non-translatable words we added as such in the query. (Airio et al. 2003.) N-gram methods were not utilized because search topics did not contain any typical OOV (out-of-vocabulary) words (proper names, geographical names or special terminology).

The translation dictionaries utilized by UTACLIR were MOT Finnish-Swedish-Finnish dictionary and MOT GlobalDix by Kielikone plc. The Finnish-Swedish dictionary includes 84 000 Finnish entries. GlobalDix is a multilingual dictionary, based on an English core. GlobalDix includes 26 000 Finnish, 30 000 French, 44 000 English and 29 000 German entries (see Kielikone plc).

Hereinafter, we call the user formulated target language query a *target query*, the user formulated source language query a *source query*, the query translated from the source query utilizing UTACLIR an *UTACLIR query*, and the Babelfish translated query a *Babelfish query*.

The queries our participants formulated were typical Google queries: short, consisting mostly of nouns. As Babelfish has been designed for translation of complete grammatical sentences, such keyword lists pose a challenge. We noticed that Babelfish sometimes translated these short queries in an unreliable way. For example, for a query *cycling Germany accommodation* Babelfish gave a translation *einen.Kreislauf.durchmachendeutschland Anpassung*. Thus, we inserted semicolons between the source words before inputting the query to Babelfish, after which Babelfish translated each query word separately: *Radfahren; Deutschland; Anpassung*.

The test system was implemented with the Java servlet technology. The first page of the interface included two boxes: one for the query in the source language, and the other for the query in the target language. When the user had typed in both queries, she / he pressed the *send* button. The queries were sent to the Java servlet which translated the source language query with our dictionary based query translation system UTACLIR (see Airio et al. 2003), and the English-German query also with Babelfish

(http://babelfish.altavista.com/). Each query in the target language was sent to Google (http://www.google.com/) separately, retrieving two or three Google result lists. (See Figure 1.)

The servlet merged the result lists, taking one from each by turn and removing duplicates. The source list for the first selection was chosen at random. There were two reasons for merging the lists. First, we thought it was fair that test persons did not know or guess the origin of documents: if they would have known, it could have affected their decisions. Second, duplicate documents would have caused problems. In the most ultimate case, the lists could share many documents. In that case, the user would have had to evaluate some documents repeatedly.

The number of documents from each list was at most nine (Finnish-Swedish and Finnish-French) or six (English-German). Thus, the total number of documents in the final result list was eighteen or less. It was less if Google retrieved less than the allowed number of documents in each case. Duplicates were not shown to the user. However, the system kept track on which systems had retrieved each document.

The merged result list, containing menus for making relevance assessments, was sent to the client-applet. When the user had made her / his assessments, she / he pressed the *send* button to effect storing of the assessments. The urls with information on their origin (retrieved by the target language query / UTACLIR translation / Babelfish translation) and relevance assessments were written into a log file.

3.6 Analysis

The research community has found that the binary relevance scale is not adequate for a real evaluation. Searchers prefer highly relevant documents over marginally relevant, which fact has not been taken into account when binary relevance scale was utilized. (See Sormunen 1994, Järvelin and Kekäläinen 2000, Voorhees 2001, Järvelin and Kekäläinen 2002, Sormunen 2002, Borlund 2000.) In addition, the relevance threshold used in traditional laboratory tests has been very low (see Sormunen 2002). Thus, the non-binary evaluation scale gives a more truthful image of the performance of various systems than the binary scale. We utilized a four-point relevance scale in our test: 0 – not relevant, 0.33 – marginally relevant, 0.66 – quite relevant, 1 – relevant.

Relavance may be interpreted to have evolving and dynamic nature. It is situational and connected with the test person's individual information need (Borlund 2000, 78-79). Regarding the personal dimensions of relevance, we decided to utilize in our

calculation the relevance assessments given by each participant, instead of calculating average scores.

Kekäläinen and Järvelin have defined the generalized precision gP in the following way:

$$gP = \Sigma\, d \in R\ r(d) / n$$

where R is a set of n documents retrieved from a database $D = \{d_1,\ldots,d_N\}$ and r(d) are relevance scores of documents, ranging from 0.0 to 1.0 (Kekäläinen and Järvelin 2002).

We restricted the number of retrieved documents nine per a query with Finnish-French and Finnish-Swedish tests and six with English-German tests. We noticed that a result list retrieved with an error-free query (no spelling errors) always included the maximum allowed number of documents. Thus, we utilized values six and nine as *n* in the formula, even if the number of documents retrieved was smaller. Thus we could avoid bias caused by erroneous queries which despite of errors could retrieve one (relevant) document, bringing the best possible precision, while a good query bringing eight relevant documents would achieve a poorer result.

The average generalized precision (with n-values nine and six) over all queries was utilized as the performance measure.

Parametric statistical tests (repeated measures) were performed in order to answer the research question two (see Black 1999, 602). The performance of the target queries, UTACLIR queries (and Babelfish queries) are the within-subject variables, while the groups (the language skills, the vocabulary familiarity and the topic familiarity) are the between-subject variables in the tests.

4 Results

4.1 Participants

Most of the Finnish-French participants had *good language skills*, while the distribution of Finnish-Swedish and English-German participants was quite even (see Table 1).

Most of the participants in each language pair had no relationship with the target language. Some studied the language at the University, some spoke the language with friends or relatives. Two of the Finnish-Swedish participants had lived in Sweden. We asked the participants whether it was easier to formulate source language or target language queries. All of our English-German participants and most of Finnish-Swedish

and Finnish-French participants thought that query formulation was easier in the source language.

The participants could give none or multiple choices for the rest of the queries. Two of Finnish-French and three of Finnish-Swedish participants mentioned that word inflection caused them difficulties in source query formulation. Some had difficulties in limiting the query.

Remembering target language words or spelling caused difficulties for all the participants except one. Two of the Finnish-French participants did not know whether to use prepositions in queries or not.

Table 1. Characteristics of the participants

	Finnish-Swedish	English-German	Finnish-French
Language skills			
moderate	10	5	3
good	8	7	9
Relationship with the target language			
no relationship	11	6	6
University studies	3	3	4
relatives / friends	2	3	2
has lived in Sweden	2	-	-
Query formulation was easier in			
source language	16	12	11
target language	2	-	1
Difficult in source query formulation			
limiting the query	5	6	3
word inflection	3	-	3
Difficult in target query formulation			
words	17	12	12
prepositions	-	-	2
Document evaluation			
scanning through	12	8	5
headings	10	6	5
checking the url	2	2	2
using “find”	3	3	-

The participants mentioned four different heuristics when evaluating the documents (see Table 1). The most common was scanning through the document. The second most popular was reading the headings. Some participants checked whether the url of the

document seemed to be reliable, and some utilized the *find* function of the browser to check whether the document included important words.

4.2 Tasks

Each participant performed four tasks. The number of Finnish-French and English-German tasks was thus 48. We had to abandon one Finnish-Swedish task because the participant did not follow the given instructions. Thus, the number of Finnish-Swedish tasks was 71. Among all the language pair groups, the alternative *a little familiar* was the most common when the topic familiarity and the vocabulary familiarity were asked (see Table 2).

Table 2. Distribution of tasks by the topic familiarity and the vocabulary familiarity

	Finnish-Swedish	English-German	Finnish-French
Topic			
1 not at all familiar	29	12	9
2 a little familiar	37	27	34
3 very familiar	5	9	5
Mean	1.7	1.7	1.9
Vocabulary			
1 not at all familiar	29	10	14
2 a little familiar	36	32	26
3 very familiar	5	6	8
Mean	1.6	1.9	1.9

The distribution of performed tasks according to the topic number was not even (see Table 3). Over half of the tasks in all the language groups were covered by three topic numbers.

Table 3. Distribution of performed tasks according to the topic number

	Number of tasks		
Topic number	Finnish-French	Finnish-Swedish	English-German
1	6	12	2
2	10	11	9
3	4	4	7
4	0	0	1
5	2	3	2
6	9	13	8
7	3	8	4
8	9	13	8
9	1	4	4
10	4	3	3
Total	48	71	48

The topics selected might have an effect on the performance: some topics are easier producing better results than the others. It might have been possible that the distribution of the topics among the language skills or the topic / the vocabulary familiarity groups was not even. In order to check for the possible effect of topic choice on the results we calculated the contingence coefficient measure between the topic number and each of our metrics. We found one correlation: in the Finnish-Swedish test the correlation between the topic number and the topic familiarity was significant at the 0.05 level - in other words, some of the topics were more familiar to most of the participants than the others. Thus, despite that single correlation, the skills or the familiarity seem not to correlate with topics, i.e. they are not properties of topics but rather properties of the relationship between topics and persons. Therefore we may safely study the effects of the skills or the familiarity on retrieval performance despite the different topic choices by test participants.

4.3 Tests

The average generalized precision of the target queries was 21.8 % in the Finnish-Swedish test, while it was 24.5 % for the UTACLIR queries (see Table 4). Differences between the results are not statistically significant at the 0.05 level. The performances differed significantly (< 0.05) in the groups according to the language skills and the vocabulary familiarity.

Table 4. Average generalized precision of the Finnish-French, Finnish-Swedish and English-German tests

	Average generalized precision %		
	target queries	UTACLIR queries	Babelfish queries
Finnish-Swedish	21.8	24.5	-
English-German	26.1	32.5	29.0
Finnish-French	33.7	16.9	-

In the English-German test, the average generalized precision of the target queries was 26.1 %, for the Babelfish queries 29.0 % and for the UTACLIR queries 32.5 % (see Table 4). Differences between the results are not statistically significant at the 0.05 level. The performances differed significantly (< 0.05) in the groups according to the topic familiarity.

In the Finnish-French test, the average generalized precision of the target queries was 33.7 % and that of the UTACLIR queries was 16.9 % (see Table 4). Differences between the results are statistically significant at the 0.05 level. The performances differed significantly (< 0.05) in the groups according to the topic familiarity and the vocabulary familiarity.

4.4 Finnish-Swedish performance

In the Finnish-Swedish test, two of the between-subject factors were significant: the language skills and the vocabulary familiarity. We do not consider here the effect of the topic familiarity, because it correlated with the topic numbers.

When the participants were classified into two groups according to the *language skills*, the average generalized precision of the UTACLIR queries in the groups was almost the same (23.5 % and 25.7 %, see Table 5). The average generalized precision of the target queries for those in the group with *moderate skills* was 12.8 %, while it was 33.5 % for the group with *good skills*. Thus, for those who belonged to the group *moderate skills*, query translation was beneficial, while it did not help the participants in the group with *good skills*.

The *vocabulary familiarity* seemed to correlate both with the results of the target queries and the UTACLIR queries: the more familiar the vocabulary, the better results (see Table 5). The groups were not even, however: the group *very familiar* with the vocabulary included only five query pairs. Thus, there may be some impact of coincidental factors.

Table 5. The average generalized precision of the Finnish-Swedish test by between-subject factors with significant effect on the performance differences

	Average generalized precision %	
	target queries	UTACLIR queries
Language skills		
moderate	12.8	23.5
good	33.5	25.7
Vocabulary familiarity		
not at all familiar	14.4	19.8
a little familiar	18.9	26.3
very familiar	74.8	35.8

We calculated the average performance for each topic. There were three topics where the UTACLIR queries outperformed the target queries. Many of the target queries for these topics were defective: the participant did not remember the Swedish words or made spelling errors. The reasons for the poor performance of the UTACLIR queries seemed to be mostly due to UTACLIR compound handling. UTACLIR translates compounds as whole, if they are included in the dictionary. It breaks untranslatable compounds into the constituents and envelopes the translations of the parts with the synonym operator, which was here converted to the OR operator. Many queries included untranslatable compounds, which caused much noise.

4.5 English-German performance

In the English-German test, both the UTACLIR queries and the Babelfish queries outperformed the target language queries (see Table 4). We can thus imply that query translation was beneficial (but not statistically significantly).

Table 6. Average generalized precision of English-German test by the between-subject factors

	Average generalized precision %		
	target queries	UTACLIR queries	Babelfish queries
Language skills			
moderate	18.3	37.8	35.3
good	31.7	28.8	24.6
Topic familiarity *			
not at all familiar	15.7	32.4	21.7
a little familiar	34.2	36.0	35.4
very familiar	16.0	22.2	19.7
Vocabulary familiarity	3.3	18.3	36.1
not at all familiar	31.1	37.7	30.4
a little familiar	38.0	28.7	10.2
very familiar			

* Significant effect at the 0.05 level on the performance differences

In the English-German test, only the *topic familiarity* had a significant effect on the performance differences between the target / UTACLIR / Babelfish queries. The participants *not at all familiar* with the topic got better result with the translated queries than with the target query, which is reasonable. Those who were *a little familiar* got as good results with all the queries. The results of the participants *very familiar* with the topic are confusing: they got poor results with the target queries. Some target queries in this group were defective (for example only one word *rauchen*, *to smoke*, was used as query word for the task number eight) and some included spelling errors, which explains the result.

We had a closer look at the results according to other two between-subject factors, the language skills and the vocabulary familiarity as well, even if they did not have any significant effect on the performance differences. The performance of the target queries was much worse than that of the UTACLIR queries and the Babelfish queries for the participants with *moderate language skills*. The participants with *good language skills* got better results with the target queries than with the UTACLIR queries and Babelfish queries, but the differences were small.

The participants who were *not at all familiar with the vocabulary* got exceptionally poor results with the target queries: many words were missing from the queries and there were a lot of spelling errors. They benefited much from query translation, because the results of the UTACLIR queries and the Babelfish queries were much better than that of the target query. The participants who were *a little familiar with the vocabulary*

got about the similar result with all the queries, while those who were *very familiar* did not benefit from query translation.

We calculated the average performances for each topic. The target queries outperformed slightly the UTACLIR and the Babelfish queries in three topics. Many English queries for these topics were imperfect: important words were missing, causing poor result for translated queries.

Babelfish outperformed UTACLIR in four topics. There was one topic where an important word was missing from the translation dictionary utilized by UTACLIR (topic 4: activist). In the other tasks, UTACLIR gave an incorrect translation from the topic point of view (for example topic 3: gay – fröhlich, bunt).

The UTACLIR queries outperformed the Babelfish queries in six topics. In most of the cases, the translation alternative given by Babelfish was not appropriate for the task (for example topic 9: alcohol - Spiritus).

4.6 Finnish-French performance

We calculated the average generalized precision for all the between-subject factors in the Finnish-French test (see Table 7).

Table 7 Average generalized precision of Finnish-French test by the between-subject factors

	Average generalized precision %	
	target queries	UTACLIR queries
Language skills		
moderate	28.2	15.7
good	50.3	20.3
Topic familiarity *		
not at all familiar	17.7	12.3
a little familiar	34.5	14.3
very familiar	57.1	43.0
Vocabulary familiarity *		
not at all familiar	22.5	2.1
a little familiar	31.9	19.6
very familiar	59.3	33.8

* Significant effect at the 0.05 level on the performance differences

The *topic familiarity* had a significant effect on the performance differences. The average performance of the target queries for participants *not at all familiar* with the topic was 17.7 %, 34.5 % for *a little familiar* and 57.1 % for *very familiar*. Thus, those who knew the topic better got better results with the target query. The average generalized precision of the UTACLIR queries was 12.3 %, 14.3 % and 43.0 %,

respectively. These results are not very reliable, however, because the group *very familiar* included only five tasks out of 48.

Also the vocabulary familiarity had a significant effect on performance differences. We can see that those *familiar with the vocabulary* got better results with the target query than those who are *a little familiar*, and the participants who are *not all familiar* got the worst results. The result of the UTACLIR queries of those who were *not at all familiar* with the vocabulary was only 2.1 %. The source queries in this group included a lot of words missing from the translation dictionary. On the other hand, the participants who were *very familiar* with the vocabulary made source queries which included words present in the dictionary, which explains their good result of the UTACLIR queries.

The result of the target queries in the group with *moderate language skills* was 28.2 %, while it was 50.3 % for those with *good language skills*. The result of the UTACLIR queries are 15.7 % and 20.3 %, respectively. Again, many of the source queries in the group with *moderate language skills* included non-translatable words.

In order to clarify the performance difference between the target language and the UTACLIR queries, we calculated the average performance for each topic, and took a closer look at those with large differences. There was one topic where the UTACLIR queries performed better than the target queries (average generalized precisions 29.3 % and 11.4 % respectively), and one where the UTACLIR queries almost achieved the level of the target queries. The good performance of UTACLIR was due to beneficial translation alternatives given by the translation dictionary. On the other hand, UTACLIR gave very poor results for the other topics. Many of the Finnish queries for these topics included words not present in the dictionary (*tupakointi – smoking, pyöräily – cycling*). Thus we can conclude that the coverage of the translation dictionary has a considerable impact on CLIR performance.

6 Discussion and conclusions

Our first research question was: what is the relative performance of the users' direct querying in the target language vs. the automatically translated target language queries. The answers were different depending on the language pair. In the *Finnish-Swedish* test, the target queries performed almost equally with the UTACLIR queries, the result of the latter being slightly better. In the *English-German* test, the UTACLIR queries attained

the best results, the second was Babelfish, while the target queries got the worst results. The differences were not statistically significant, however. In the *Finnish-French* tests, the target queries performed twice as well as the UTACLIR queries, and the differences were statistically significant. The Finnish-French translation dictionary had shortcomings which affected the results. We found that the quality of the translation dictionary is very important: a defective dictionary does not help even those with not so good language skills.

In conclusion, when the result of the Finnish-French test is downplayed because of the poor dictionary, query translation compared with target language queries seemed to perform much better in our test than it has performed in previous laboratory tests. The reason might be that in the laboratory tests the target language queries are worded on the bases of the topics, which have been formulated by native language speakers. Thus, the queries include all the important words with no spelling errors. On the other hand, the target queries formulated by test persons often are defective: they do not cover all the perspectives of the topic, they include spelling errors, they are too specific or too loose. Formulating a good query is not easy, especially in a foreign language. Thus, our test gives a more truthful image of the benefits of query translation for real users than the laboratory tests do.

Our second research question was whether the language skills and the topic familiarity / the vocabulary familiarity have any effect on the performance differences between the target / UTACLIR / Babelfish queries. In the Finnish-Swedish test we found that the better language skills the user had and the more familiar she was with the vocabulary, the better were the results of the direct queries, while the results of the translated queries were approximately the same between the groups. The results of Finnish-French and English-German tests were parallel, even though the corroborating findings were not statistically significant.

Thus, we can conclude that our assumption was right: the benefits of query translation in the Web seem to depend on users' language skills and activity. We did not have participants with poor language skills in our test, but obviously that kind of users would benefit even more from query translation than those with moderate language skills. However, they would need quality translations of the documents found.

The next issue to study would be whether multi-lingual query translation (one source language, many target languages) is advantageous for Web users.

Acknowledgements

Research funded, in part, by Academy of Finland project no. 1209960.

The author wishes to thank Prof. Jaana Kekäläinen and Prof. Kalervo Järvelin for their comments in preparing this article.

ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright (c) 1989-1992 Atro Voutilainen and Juha Heikkilä.
FINTWOL (Morphological Description of Finnish): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1993.
GERTWOL (Morphological Transducer Lexicon Description of German): Copyright (c) 1997 Kimmo Koskenniemi and Lingsoft plc.
TWOL-R (Run-time Two-Level Program): Copyright (c) Kimmo Koskenniemi and Lingsoft plc. 1983-1992.
GlobalDix Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone plc, Finland.
MOT Dictionary Software was used for automatic word-by-word translations. Copyright (c) 1998 Kielikone plc, Finland.

Appendix 1

The start questionnaire

The year of graduation:

The last report grade in the target language (5-10):

The high school graduation exam grade in the target language (a, b, c, m, e, l):

Have you practiced your skills in the target language by reading and/or listening over the past two years?

1 = not at all
2 = a little (a couple of times in a year)
3 = monthly
4 = much (at least weekly)

Have you practiced your skills in the target language by speaking and/or writing over the past two years?

1 = not at all
2 = a little (a couple of times in a year)
3 = monthly
4 = much (at least weekly)

Do you have any other relationship or activity regarding the target language?

Appendix 2

The topic questionnaire

Was the topic familiar to you?

1 = not at all
2 = a little
3 = very familiar

Was the target language vocabulary (concerning the topic) familiar to you?

1 = not at all
2 = a little
3 = very familiar

Appendix 3

The closing questionnaire

Which did you find easier, formulating the source language or the target language query?

What was difficult when formulating the queries?

Which document representations and/or parts did you use to assess ...? [snippets and browsing full text].? What was difficult?

Appendix 4

English-German topics

1. You have a German friend, who has lived in Finland for three years. He and his family have converted to Islam a year ago. They have decided to move back to Germany. There are two school-aged children in the family. Your friend wants them to attend an Islamic school in Germany. On the other hand, he has heard that all the German citizens, no matter their ethnic and religious background, have to attend basic studies in German state schools. You promised your friend to clarify this by searching information in the Internet.
2. You have friends who are very interested in German culture, language and food. You are going to organize a German evening for them: play German music, show German movies and offer German food and bier. The only German food speciality you remember is sour cabbage. Try to find German recipes in the Internet, sour cabbage recipes and also other ones. The recipes must be written in German, naturally!
3. You have heard that one of your friends is a gay, but he has told you nothing about this himself. During the last few months this friend has considered the possibility to move to Germany. Now you are curious to know what has been written in the Internet in German about gay marriages and German legislation concerning gay marriages.
4. You are interested in the animal activism: you always read the news concerning activists' measures, especially against fur farms. Now you wonder what the situation in Germany is: are there animal activists in Germany and do they come against furs? You want to find documents in the Internet describing measures taken by animal right activists against fur farming, fur shops and people wearing furs.
5. You have chosen a new minor subject at the university: social work. Now it is time to write your first essay. Your subject is children taken in custody because of drug use of the parents. You want to make comparative study between cases in Finland and Germany. You try to find documents in the Internet on that kind of cases in Germany.
6. You and your friends are planning a cycling tour to Germany next summer. None of you has visited Germany before, but all of you have some German language skills. Now you should clarify whether there are any recommended cycling tours in Germany, and what kind of accommodation is available. You make an Internet search to find out these things.
7. You are interested in all kinds of special diets, and you always read texts concerning them. You don't adhere to any diet yourself, but you rather are willing to question the rationality of them. In Finland diets of low carbohydrate are quite favourable nowadays. You want to find out whether they are popular in Germany as well. You want to know what has been written about those diets and their slimming effect in the Internet in German.
8. Many of your friends smoke. They are not very rapturous about the possible new law whose purpose is to restrict smoking in public places, such as restaurants, cafes and bars. You ponder what might be the situation in Germany: are there any laws restricting smoking? What do people think about smoking restrictions? In order find out this you make an Internet search trying to discover German discussions and opinions about smoking restrictions.
9. You have followed the conversation about legalizing drugs in Finland. There are opinions that alcohol is much more dangerous than cannabis and other mild drugs, and that smoking is as addictive as using drugs. Now you want to know what kind of opinions there are in Germany, and you use the Internet for clarifying this.
10. You have taken journalist studies as a part of your degree. Now you are participating in a course where you should write an essay about texts in the Internet: what kind of texts are there about the subject you have selected. In addition, you have to compare Finnish texts with texts in some other language. You select the mental health problems among young people as your subject, and German as the other language. Now you should find German texts about mental health problems among young people in the Internet

References

Airio E, Keskustalo H, Hedlund T and Pirkola A (2003) UTACLIR@CLEF2002 – Bilingual and multilingual runs with a unified process. In Peters C, Braschler M, Gonzalo J and Kluck M, (eds.), Advances in cross-language information retrieval. Results of the cross-language evaluation forum - CLEF 2002, Lecture Notes in Computer Science 2785, Springer, pp. 91–100

Bian, G.-W., & Chen, H-H. (2000). Cross-language information access to multilingual collections on the internet. Journal of the American Society for Information Science 51(3), pp. 281–296.

Black, T.R. (1999). Doing quantitative research. in the social sciences. An integrated approach to research design, measurement and statistics. SAGE Publications.

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

Capstick, J., Diagne, A.K., Erbach, G. Uszkoreit, H., Leisenberg, A. & Leisenberg, M. (2000) A System for supporting cross-lingual information retrieval. Information Processing and Management 36 (2), pp. 275-289.

Gulli, A., Signorini, A. (2005). The Indexable Web is More than 11.5 Billion Pages. In Special Interest Tracks & Posters Publications, The 14th International World Wide Web Conference 2005, pp. 902-903. Available at http://www.www2005.org/cdrom/docs/p902.pdf.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. ACM SIGIR 1993, Pittsburgh, PA.

iCLEF interactive track for the Cross-Language Evaluation Forum. Retrieved May 2006. Available from <http://nlp.uned.es/iCLEF/index.htm>.

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In: Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '00), Athens, Greece, July 24-28, 2000. New York, NY: ACM Press, pp. 41-48.

Kekäläinen, J. & Järvelin, K. (2002). Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology 53(13): 1120-1129.

Kielikone plc. http://www.kielikone.fi/en/

Oard, D.W., Gonzalo, J., Sanderson, M., López-Ostenero, F., Wang, J. (2004). Interactive Cross-Language Document Selection. Information Retrieval 7(1-2), pp. 205-228.

Ogden, W., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H., et al. (1999). Keizai: An interactive cross-language text retrieval system. Paper presented at the Machine Translation Summit VII, Workshop on Machine Translation for Cross-Language Information Retrieval, Singapore, PRC.

Ogden, W.C., & Davis, M.W. (2000). Improving cross-language text retrieval with human interactions. Proceedings of the Hawaii International Conference on System Science (HICSS-33), Vol. 3.

Over, P. (1998). The TREC interactive track: an overview. Retrieved October 2005. Available from <http://www.itl.nist.gov/iaui/894.02/works/presentations/dublin98/sld001.htm>

Penas, A., Gonzalo, J., & Verdejo, F. (2001). Cross-language information access through phrase browsing. Paper presented at the 6th International Conference of Natural Language for Information Systems (NLDB'01), Madrid, Spain.

Petrelli, D., Levin, S., Beaulieu, M. & Sanderson, M. (2006). Which user interaction for cross-language information retrieval? Design issues and reflections. Journal of the American Society for Information Science and Technology 57(5) pp. 709-722.

Sormunen, E. (1994). Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa Licentiate thesis at University of Tampere.

Sormunen, E. (2002). Liberal relevance criteria of TREC - counting on negligible documents? In: Beaulieu, M. et al. (Eds): Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval. August 11-15, 2002, Tampere, Finland. Special Issue of SIGIR Forum 36:324-330.

TREC-2002 interactive track home page. Retrieved October 2005. Available from <http://trec.nist.gov/data/t11_interactive/t11i.html>.

Text Retrieval Conference (TREC). (2005) NIST – National Institute of Standards and Technology. Retrieved September 2005. Available from <http://trec.nist.gov/>.

Voorhees, E. Evaluation by highly relevant documents. (2001). In: Croft, W.B. et al. (eds.). Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New Orleans, September 2001). ACM Press, New York, 74-72.