Which User Interaction for Cross-Language IR?

Design Issues and Reflections

Daniela Petrelli Steve Levin Micheline Beaulieu Mark Sanderson
Department of Information Studies – University of Sheffield
Regent Court – 211 Portobello Street
S1 4DP –Sheffield - UK
d.petrelli@shef.ac.uk

ABSTRACT

A novel and complex form of information access is cross-language information retrieval: searching for texts written in foreign languages based on native language queries. Although the underlying technology for achieving such a search is relatively well understood, the appropriate interface design is not. This paper presents three user evaluations done during the iterative design of Clarity, a cross-language retrieval system for rare languages, and shows how the user interaction design evolved depending on the results of the usability tests. The first test was instrumental to identify weaknesses in both functionalities and interface; the second was run to determine if query translation should be shown or not; the final was a global assessment and focussed on user satisfaction criteria. Lessons were learned at every stage of the process leading to a much more informed view of what a cross-language retrieval system should offer to users.

1. Introduction

Information is produced daily in the most diverse languages. This abundance of information, possibly relevant to communities other than the one that produced it, stimulated research since the early seventies when experiments for retrieving information across languages were first initiated (Salton, 1973). A whole branch of Information Retrieval (IR) has been devoted to overcome language boundaries: *Cross language information retrieval* (CLIR) is the retrieval of information written in one language based on a query expressed in another, e.g. typing a query in English to retrieve documents written in Finnish. For such a process to succeed, translation of the user query written in the *source language* (e.g. English) or of the documents written in the *target language* (e.g. Finnish) must occur.

During the nineties much effort was spent experimenting with different techniques, and a collective effort of IR researchers (TREC; CLEF; NTCIR) produced systems able to retrieve effectively (Ballesteros & Croft 1998; McCarley, 1999; Xu & Weischedel, 2000). However, the effort was mainly directed toward retrieval functionality and effectiveness; little attention was paid to potential utility of CLIR and users were rarely involved. Unverified assumptions were made such as that users would only have limited knowledge of the target language (if any) thus requiring some kind of translation at display time in order for the user to detect relevant documents (Oard, 1997). Conversely a more holistic study (Capstick et al, 1998) suggested that *polyglots* (people who speak more than one language) were potential users of CLIR systems. This result opened a new perspective on CLIR users and uses, and would affect the way CLIR systems are designed with specific reference to the user interface and interaction. However no other studies followed that initial investigation, thus previous knowledge was of limited help to us in Spring 2001 when starting the design of our CLIR system, Clarity[1]. The final prototype allows users to perform multilanguage search for so called *low-density languages*, i.e. languages for which electronic resources are limited. Besides English, the other languages Clarity encompasses are Finnish, Swedish, Latvian, and

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

Lithuanian. Clarity is efficient and effective and has been well received in the final user evaluation performed in December 2003. Though, the final interaction design slowly emerged during an iterative evaluation-redesign cycle. A user-centred approach was adopted. An extensive user study suggested the main directions for the initial design (Petrelli et al. 2002). Notably the study pointed out a conflict between user requirements and good practice in interactive CLIR. Two different interfaces were then developed that respected the two positions and were tested in July 2002. The result was inconclusive so a redesign took place and a new evaluation was performed. This test showed a tension between the most effective interaction (based on IR literature) and the most appreciated interface design (based on user requirements). Empirical evidence supported the decision of adopting the more effective but less favourable interaction, and a redesign took place, which was tested in December 2003. Results showed casual users could retrieve documents effectively and had positive opinions of the system used.

This paper describes the evolution of the interaction design through the empirical results of the three evaluations. The first evaluation is fully reported elsewhere (Petrelli et al., in press) and therefore it is only summarized in section 2. The second evaluation is then fully discussed in terms of experimental conditions, results, and derived suggestions (section 3). The final layout and usability test follow in section 4. Interface design issue (section 5) and reflections of the use of multilingual IR (section 6) conclude the paper.

2. First Discovery of User-CLIR Interaction

Clarity's first design was based on CLIR and IR literature and required users to supervise query translation. By explicitly involving users we intended to support their understanding of CLIR mechanisms and provide full control over the system. However the advantage was only hypothetical: a user study to elicit actual needs (Petrelli et al., 2002) contradicted the main assumption as the core decision of letting the user supervise the query translation was disliked in favour of a simpler layout. This result called for a different interaction and a comparative user test was undertaken to empirically investigate the two approaches:

- **Supervised mode (SM):** derives from the CLIR/IR literature and requires the user to input the query first, then query translation is shown for user verification and/or modification, and finally the system searches, Figure 1a;
- **Delegated mode (DM):** derives from the user requirements and entails the user to only input the query, then the system translates the query and searches without any user intervention, Figure 1b.

The two interactions corresponded to two user interfaces that were kept as similar as possible to avoid bias. The main difference was on performing two steps (translate and search) or a single one to get the results; indeed even in the DM users could see the query translation if "see query translation" was selected. In both layouts to modify their query users had to enter a new query in the box.

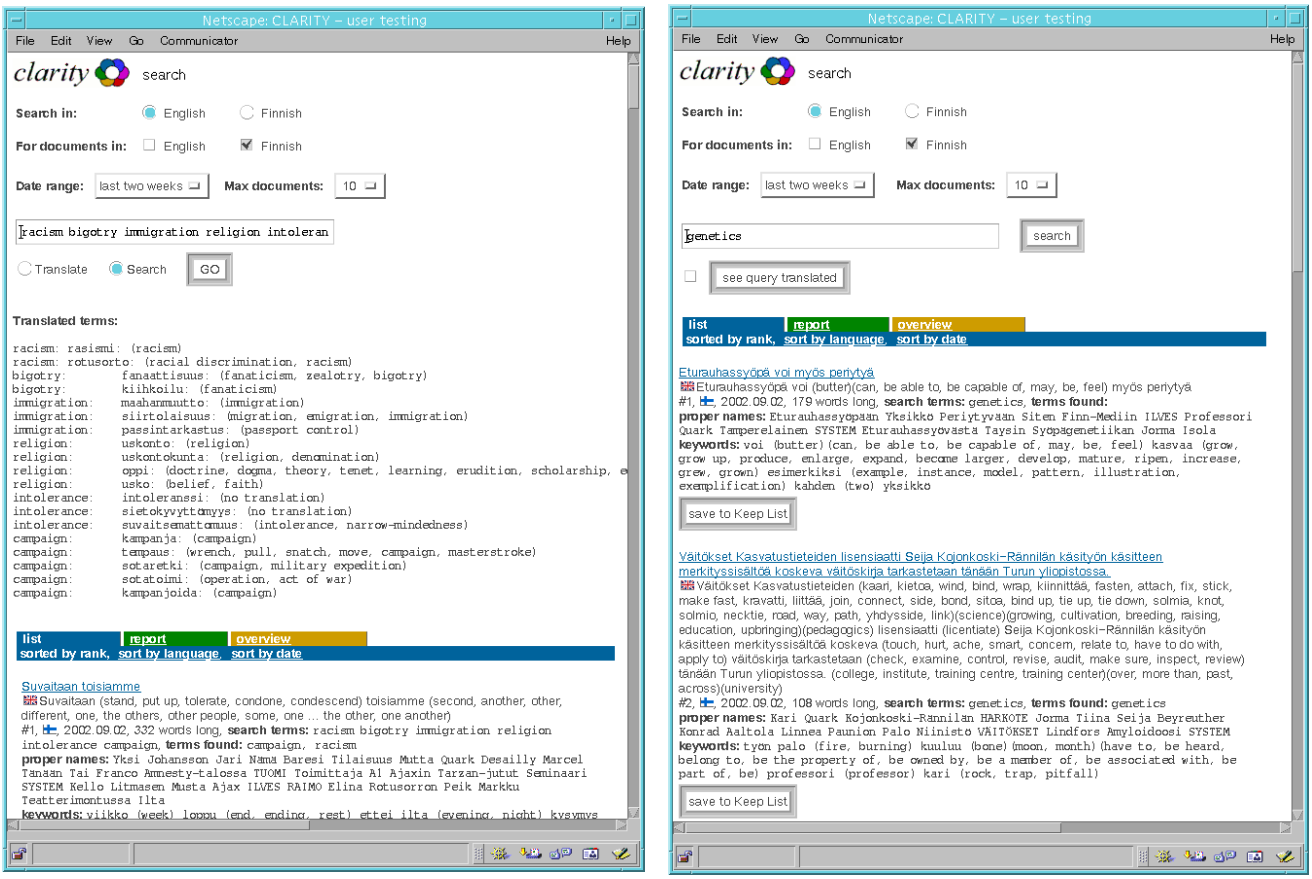

Figure 1. The Clarity interface as tested in the first usability test: a) user-supervised; b) delegated mode.

Six monolingual (English) users participated in the experiment and searched a Finnish collection. The interactive CLEF experimental framework (Gonzalo & Oard, 2002) was followed, but additional measurements (both objective and subjective) were taken. Results were many and affected the redesign of the system architecture, features, and interaction (Petrelli et al., in press). Here only the main points relevant for the actual discussion are reported.

The effectiveness of each layout was assessed by average precision and recall measured at display time, i.e. before the users bookmarked the relevant documents. The low values reported in Table 1 are due to a task for which none of the subjects was able to retrieve any relevant document.

	Precision	**Recall**
Supervised	0.18	0.22
Delegated	0.161	0.123

Table 1. P & R for the two systems measured at display time.

An overwhelming preference for the Delegated mode (70%) over the Supervised one (15%) emerged from the questionnaires (Table 2). Almost all subjects preferred the interface that hides the translation (Fig. 1b) even if the difference between the two was rated as minimal.

	Supervised	**Delegated**	**No difference**
Easier to learn	0	70%	30%
Easier to use	15%	70%	15%
Best overall	15%	70%	15%

Table 2. User preferences as measured in the first evaluation.

Proper names were widely used by users (50% of the subjects) but badly managed by the system. Some names were in the dictionary (e.g. Europe) thus were translated, others where not (e.g. Alzheimer), and others were wrongly translated (e.g. Bobby Sands a famous hunger striker was translated into the Finnish equivalent of "policeman beach"). A new feature was introduced to mark terms that must not be translated. In the new prototype the query "computer @computer" searched the Finnish database for "tietokone computer" thus assuring the retrieval of documents where only the English word "computer" occurs.

A second important result for CLIR relates to the visualization of word translations. In the tested prototype all the possible translations of all the senses of polysemic words were displayed. Figure 1a shows an example of query translation: each word in the query was translated into many senses and each of those was back translated into English, again using all possible senses. Figure 1b better shows the effect of polysemic words in output: document title and keywords were translated using up to 11 terms. The inclusion of all the senses made the search inefficient and users confused when, for example, "golf pitch" was proposed as translation of "green". Indeed highly ambiguous words were critical to users that attempted to focus the query before the search was issued: A user was observed typing "green power" at first and ending searching using "wind turbine" because of the high ambiguity of the two more generic terms. Showing the query translation affected the search strategy as it encouraged revising and rethinking of the query. This explorative behaviour could potentially make the search session more effective by retrieving highly specific documents, but could also negatively affect the search as more generic but still relevant content could be discarded.

To minimize the negative effect of polysemic words, in the second prototype the number of translations was reduced to the most common senses. This choice simplified the query translation step that offered few check boxes to users to deselect unwanted senses, as in Figure 3. This solution automatically simplified the result display as titles and keywords were translated using a similar mechanism.

3. Investigating User-CLIR Interaction

Unfortunately, the data collected during the first test was not consistent enough to let the Clarity designers decide on the final layout and interaction. Thus a further user evaluation was conducted in Summer 2003. For this a new prototype was developed in terms of system architecture, functionalities, and interface design. Specifically documents were retrieved 10 at the time; word translation was limited to the most common senses; search for phrases and translation-bypass were both supported. In addition Clarity second prototype could retrieve documents written in English, Finnish and Swedish independently as well as simultaneously. The interface layout (Figure 2 and 3) was greatly simplified and included some new features as described in the next section.

3.1. Experimental Conditions

The second test was set up to finally determine which interaction had to be preferred, i.e. if CLIR should require the user supervision or could be fully delegated. The two conditions were contrasted and, as previously, the two interactions corresponded to two different user interface layouts, kept as similar as possible to avoid bias. Using the Delegated mode (DM, Figure 2), the user simply enters their query, clicks the 'Search' button and the results are then displayed. There is no user intervention during the query translation process. To modify their query, the user must re-enter the query in the box.

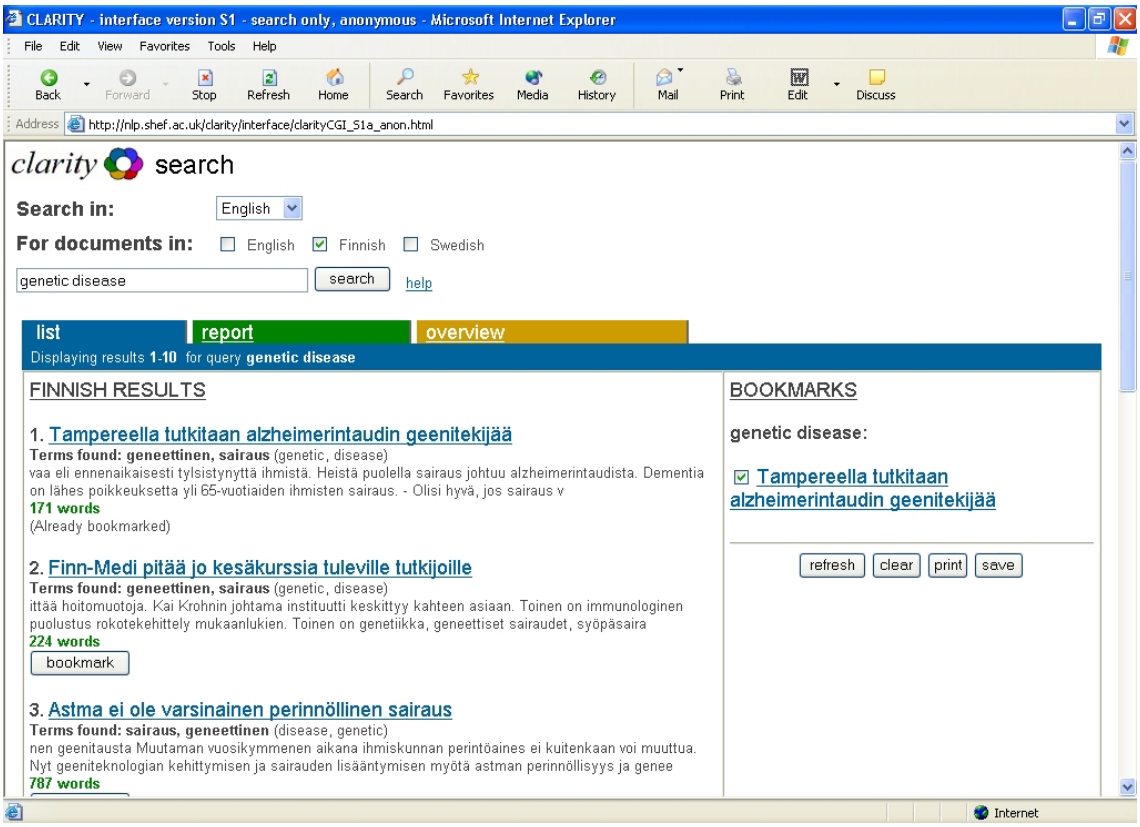

Figure 2 The Delegated layout as tested in the second evaluation (the first document has been bookmarked).

In the Supervised mode (SM), the user enters their query and clicks the 'Translate' button, they are then presented with another screen which lists the translations for each query term along with their appropriate back-translations in parentheses, as shown in Figure 3a. The translations are arranged in columns, with check boxes next to each translation to de-select unwanted senses; the user can also insert a new query, should they wish to, and ask for a new translation. Once the query translation satisfies the user, they click the 'Search' button and the results are displayed beneath the translations, as in Figure 3b.

In both interfaces, the user cumulates relevant documents in the right hand pane by clicking the 'Bookmark' button beneath each result. The pane displays titles of the bookmarked documents, which serve as links; documents can be removed by unchecking the adjacent check box and clicking 'Refresh' (Figure 2).

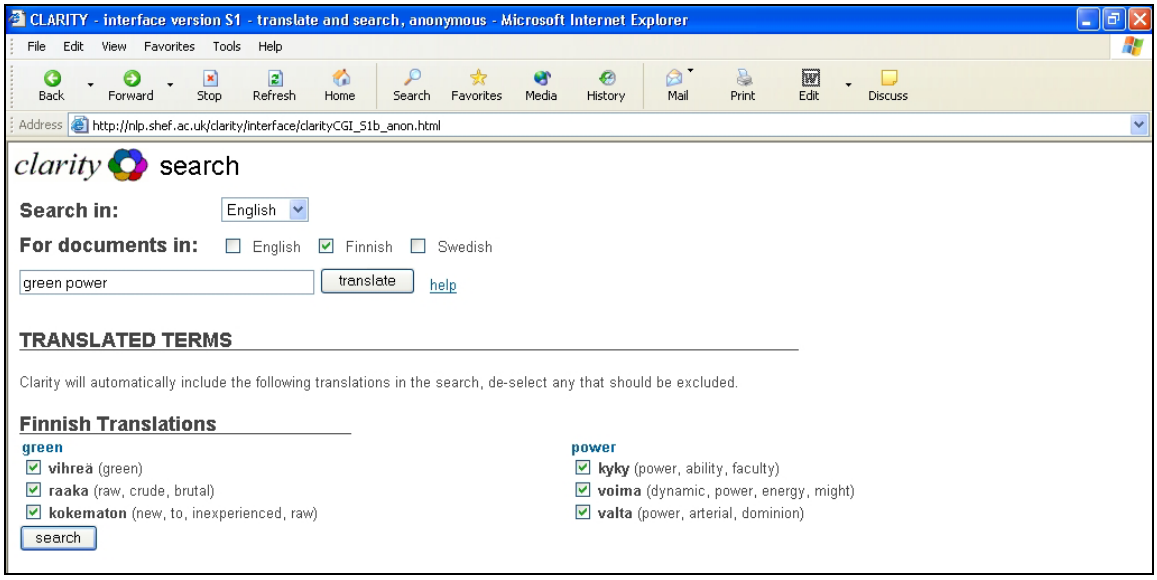

Figure 3a The query translation in the Supervised layout as tested in the second evaluation.

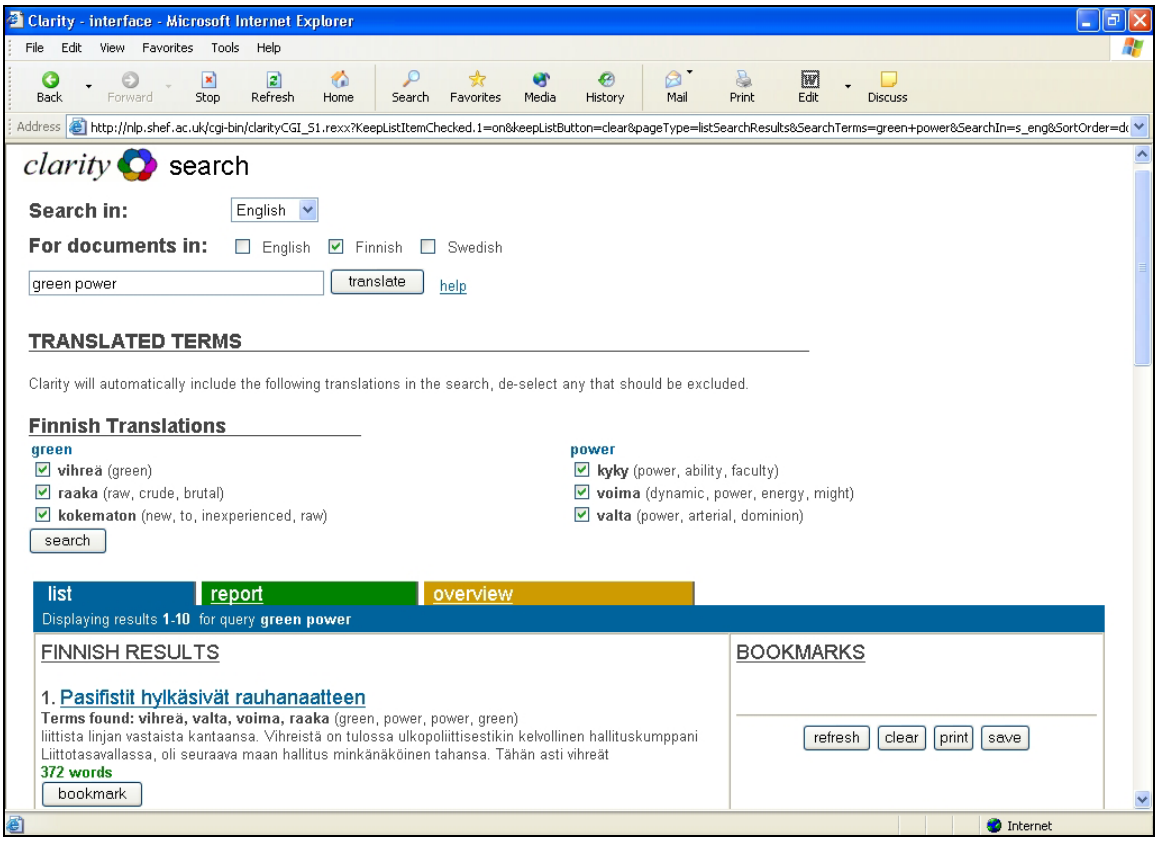

Figure 3b The results displayed in the Supervised layout as tested in the second evaluation.

3.2. Participants

For this evaluation polyglots were recruited as they seemed to be the more likely user group for CLIR (Petrelli et al., 2002). A total of sixteen participants were involved, comprising of both native Finnish and native Swedish speakers who also spoke fluent English. This enabled four different query/document language pairs to be tested (Finnish to English, English to Finnish, Swedish to English, and English to Swedish). Participants were divided into groups of four, and each group

tested a specific language pair. A participant's group allocation depended on their linguistic capabilities. Thus native Finnish speakers living in the UK were required to use English as query language and retrieve Finnish documents; the assumption was that those people would have proficient English that could resemble native knowledge[2]. Similarly native Swedish speakers living in the UK used English to retrieve Swedish documents. This first part of the test took place at The University of Sheffield, UK. A further eight participants were tested outside the UK - four at The University of Tampere, Finland (searching from Finnish to English), and four at SICS, Stockholm, Sweden (searching from Swedish to English). The experimental conditions were replicated as precisely as possible at each site to avoid introducing extraneous variables. Participants were either students or academic professionals, and were paid £15 / €20 for participating.

3.3. Procedure

The whole experiment was scheduled to last 90 minutes. At arrival, participants received a written briefing on the purpose and procedure of the test. A first questionnaire to collect personal information (e.g. education, age, languages known) and attitude towards information retrieval was filled in.

Participants were then asked to complete two retrieval tasks, one for each system layout (i.e. a within-subject experimental design was used). The tasks selected were those for which most relevant documents were retrieved in the previous test. It is worth noting that no training was offered as we were interested in observing how users first approach a CLIR system and training might hide interesting phenomena (Petrelli et al., in press).

To avoid bias, the order in which the systems were used and the task-system allocations were counterbalanced, i.e. every possible combination was tested equally. Each participant tested Clarity individually and was observed by an experimenter who noted problems and interesting interactions for the follow-up interview. Simulated tasks (Borlund 2000) were used and participants were invited to find as much information about a topic as possible and bookmark any relevant documents retrieved. The search was scheduled to last twenty minutes, but participants were informed that they could stop whenever they wished.

After completing the first task, users were asked to fill in two questionnaires, one about their familiarity with the searched topic, the other addressing user satisfaction. This was based on QUIS (Chin, Diehl & Norman, 1998) and asked participants to rate individual aspects of the system including layout, terminology, learning effort, and system capabilities. Participants were also invited to list the most positive and negative aspects in the interaction. The second task was conducted in exactly the same way but using a different system, following which the same questionnaires were completed.

The last questionnaire addressed systems comparison and asked to rate how users found the two systems, which one was easier to learn, which easier to use and which one they liked best overall.

Finally participants were interviewed. A semi-structured approach was adopted to collect participants overall reaction to the two systems as well as specific comments.

3.4. Data Collection and Analysis

The data collected was rich for both subjective and objective measures. As discussed above, several questionnaires were filled at different points in the evaluation and a final interview contributed to precisely define users' opinion.

---

[2] An initial attempt to recruit English native speakers with proficient Finnish as second language was not successful.

Objective measures were automatically recorded and time-stamped by the system: information such as queries issued, translations select/de-select, results returned, documents opened, documents bookmarked was recorded in log files. The participant's onscreen activity was also recorded using video capture software.

This rich collection of data produced a large set of results that are only summarized in this paper (details in Levin & Petrelli, 2003).

Participants were quite homogeneous with most people using Web search engines several times a week, and searching in languages other than their native one several times a week. They never use commercial search engines, and all except four had no training in information retrieval. Nevertheless all of them felt confident when searching.

To assess the overall effectiveness of each layout in supporting query formulation, average precision and recall measures were calculated. The measurement took place at display time, before the users bookmarked the relevant documents; this was done to avoid that differences in participants' judgement affect the objective values. Table 3 reports the results. Although SM performed better than DM in terms of precision and recall, the differences were minimal and not statistically significant when a paired-samples t-test was applied. However, such small differences are still meaningful as it corresponds to at least one more relevant document being retrieved out of 12-17 available in the collection, that is to say a 6 to 8% increase.

	Precision	**Recall**
Supervised	0.206	0.473
Delegated	0.167	0.418

Table 3. Precision and Recall.

Large differences emerged from language to language. The best performer was searching from English to Finnish that increased recall from 0.22 in the first evaluation to 0.838 now. This may be attributed to the improvements made to the system as a result of the first test. The worst language pair was Finnish to English since for one task none of the users was able to retrieve any relevant document. This negatively affected the overall system effectiveness as it produced precision and recall values of 0. However, it did not affect the comparison as the counterbalancing equally distributed the effect over the two conditions.

These measures only account for system performance, but equally important is to consider users' satisfaction, their thoughts and feelings, and their overall preferences. Table 4 below summarise the users preference respect to the two layouts. The Delegated mode (DM) is still the preferred one, but the divide is far smaller than the one recorded in the first evaluation (see Table 2). Indeed users feel that the system which requires supervision (SM) is no more difficult to learn nor more difficult to use than the system which does not (DM) (interviews explain why, see next session).

	Supervised	**Delegated**	**No difference**
Easier to learn	25%	31%	44%
Easier to use	44%	50%	6%
Best overall	37.5%	50%	12.5%

Table 4. User preference as recorded in the second test.

As a further support, results from the usability satisfaction questionnaire showed little difference between the users' opinions of the two systems in most areas. Notably participants rated the difficulty of using both systems as identical, thought a wider range of responses were given for DM. From this data it seems that the effort spent in improving functionalities and interface after the first evaluation was definitely worth doing.

3.5. User Comments and Experimenter's Observations

The interviews run after the test gave more insight and supported a better interpretation of the questionnaire results. Participants who favoured DM commented that it was quicker and required less effort, e.g. 'there are no extra buttons or steps, you just use it', 'it was easier… quicker'. This suggests that rather than disliking SM because it was difficult or less effective, it was generally disliked because it slowed down the searching process. Two of the three participants that in the interview gave negative feedback commented on the select/de-select feature saying it was a time consuming and not-needed step. Here a further insight: 'it was quite easy and straightforward, […] but I possibly didn't use the check boxes all the time' suggests the participant felt somehow obliged to modify the translation. A suggestion for a new layout comes from another comment: '[SM] should always assume you want all the results, then if you wish to exclude translations you can do that later'.

Although the majority preferred DM, several liked SM instead. Participants who preferred SM commented on the usefulness of seeing, checking, and updating the query translation and more in general that 'you could work with translated terms in SM… this gave you a more dynamic view of the system'. However, the comment 'it's more practical to be able to verify the translations, but it's no use if the system doesn't translate properly' highlights the fact that some users may have judged SM unnecessarily negatively as they could actually see the translations, whereas erroneous translations were not visible in the DM system. This is a crucial point for CLIR: observations of the actual interaction showed users struggling and getting frustrated when words were not translated as they expected, e.g. the Finnish "rasismi" was not translated into "racism" as the only proposed translation was "racialism" considered inappropriate by the participant. This was not an isolated case: another participant searching Swedish commented on the poor dictionary as the only translation for "discrimination" was "keep apart" and only documents about apartheid in Palestine, South Africa, and Yugoslavia were retrieved.

The solution would have been for the user to correct the system's translation or to use a synonym that would generate another translation. However users may not be skilled enough to develop different searching strategies[3]. Still synonyms might not be generated, e.g. proper names of countries: then the only chance is to directly rectify system translation.

Another comment on SM says that 'through the translations you got inspired to use other words and you saw other possibilities regarding re-formulation'. Seeing the query translation (SM) has a twofold positive effect: on one side it allows to improve the translation, on the other side it prompts users to rethink their original query. Several participants were observed deselecting the Finnish "sotaretki" as translation of "campaign" meaning military campaign while they were concerned with anti-racism campaign. It should be noted that the back-translation into English did not help in this case as it was again "campaign" and the fact that the users could rectify the system was because of their linguistic knowledge. This is a clear example of the fact that polyglots are the ideal users able to correct the system if this is needed.

In the first experiment seeing unexpected and unsatisfying translations induced users in changing the query before the search was issued. We then checked if that behaviour was still present with the new interaction. Only one participant systematically changed the query before searching by adding new terms. In other few cases the query was changed when the translations was not satisfying, e.g. as in the "racialism" example above. It can be assumed than that the new layout did not stimulate that potentially negative behaviour thus search effectiveness was not hampered.

---

[3] The experimenter suggested using a synonym when a participant explicitly complained about the translation and wanted to rectify the problem.

Finally we wanted to investigate if the interaction mode affected the engagement with the search task. As measure of engagement the number of queries issued was initially considered. Though the mean of queries issued in DM is higher than in SM the difference is not statistically significant. However SM offers the user another way to interact other than inputting a new query. Thus for DM only the number of queries was used as measure of engagement, while SM measure includes both the number of queries and the number of deselected terms. A paired-samples t-test was conducted: There was a statistically significant increase in the engagement from DM (M=6.23, SD=3.44) to SM (M=9.62, SD=5.05), $t(12)=-4.58$, $p<.001$. Indeed the possibility of deselecting terms was central as all the users deselected at least one sense (and up to 6) from those offered by SM. The number of de-selection depends on the words used in the context of the search task.

4. Steps towards a Usable CLIR

The second experiment was run to empirically determine which interaction should be preferred for CLIR. As discussed above, a conflict between objective (precision and recall) and subjective results (questionnaires) was discovered: the most effective interaction (SM) was not the most preferred (DM). Consistently with the initial study (Petrelli et al 2002), user favoured the simplest interaction (DM) but the difference in participants' opinion was small. Similarly the effectiveness of the system in SM was only marginally superior to DM and not statistically significant. Therefore, the final Clarity redesign could take any direction. The insight gained with the interviews was of paramount importance in deciding the final layout and interaction. The final prototype automatically translates and searches; then the query translation is displayed on top of the result list, as in Figure 4. This solution has the advantage of keeping the search task a single action but shows the query translation step at the same time thus allowing for translation supervision if the user intervention is appropriate. The buttons labels changed accordingly (compare Figure 3 and Figure 4). The result display was largely kept the same; only the title translation was added (below the original title); this was considered important as the most of the users judged the document relevance just browsing thought the result list and did not enter the document page.

4.1. Testing Clarity Usability

The final prototype evaluation occurred at the end of the project in December 2003 and January 2004. Conversely from the previous two that were formative, this was a summative evaluation (Preece et al., 2002) of the work done over three years and aimed at assessing the value of the system as a whole from the user point of view. Personal opinions (collected via questionnaires and interviews) were therefore considered the main source of data, though log files were recorded asin the previous evaluation.

The final prototype was tested in all its aspects (all the languages and all the features); the system was physically distributed among UK, Finland, and Sweden and was accessed as a Web service (Dedmetriou et al, submitted). The evaluation took place at the user premises: 8 participants tested Clarity at Alma Media in Tampere, Finland; while 3 were at BBC Monitoring in Reading, UK. Participants were professionals likely to use CLIR technology in the future, i.e. journalists, librarians, information professionals, translators. All but two BBC employers were polyglots in English and Finnish, had fair/good knowledge of Swedish, but no understanding of Latvian. Participants were invited to search for predefined topics in both cross language (Finnish to English) and multi-language conditions (English to Finnish and Swedish), and to search for a topic of their own choice[4] (English to Latvian).

[4] A fourth task was to brows a cross-language concept hierarchy; however interesting this feature is not discussed in this paper.

Tasks and questionnaires developed for the previous evaluation were used. Participants were requested to verbalize their thought; this way we could check if our design choices were straightforward and if the interaction was effective.

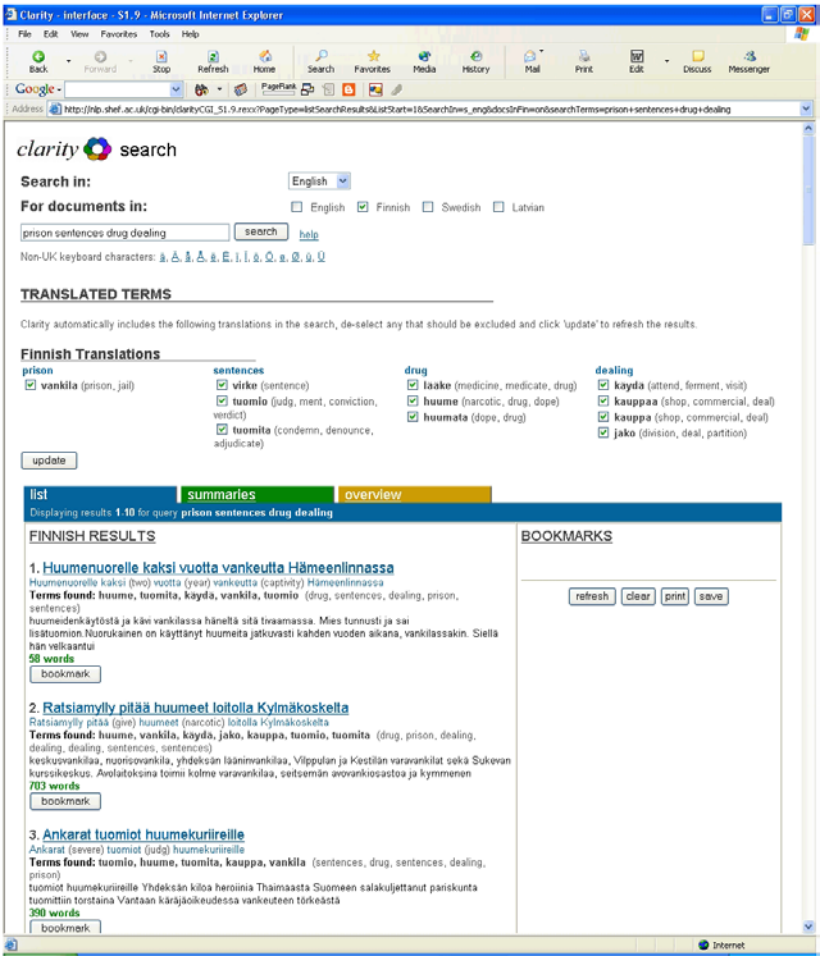

Figure. 4 Clarity final layout.

Results showed that the final system was robust, fast, accurate, easy and appealing to casual users (for details see Levin & Petrelli, 2004). Comments were extremely positive and critiques were limited to minor problems, e.g. keep the translation selection from one turn to the other; avoid automatic scrolling to the page top when bookmarking. Clarity demonstrated to be as effective with previously tested language paths as with the new one, i.e. English queries issued to retrieve documents in Latvian. Topics chosen by participants included: the Eurovision Song Contest, the restoration of Riga's Opera House, the status of Russians in Latvia, and Latvian foreign policy. All participants thought the system had retrieved documents relevant to their query and felt the translated titles and translations of terms found in the documents were helpful enough to be able to judge whether a document was relevant or not. However, one participant remarked that it did take time for them to understand the translated titles, whilst another stated that they were not always meaningful. This is due to the word-by-word translation adopted that does not consider the phrase context.

Participants were also asked if they thought searching documents in languages they do not know could be useful and why. Comments were positive; a free-lance journalist said "I could find new

interesting stories or check out something I have heard about … I would then use a dictionary or ask a friend for translation"; an information professional said "if I have done an extensive research [in English, Finnish, Swedish] I might have a look on how is the situation in Latvia … if there is something interesting then I could pass it to a translator"; a librarian said "I already help customers searching in languages I do not know … and it is difficult and frustrating… [CLIR] would help me a lot".

Searching in an unknown language was also the condition for two participants at BBC Monitoring who had no knowledge of Finnish, Swedish, or Latvian. One claimed the limited amount of translation given (title and 'terms found' only) made it too difficult to judge which results were relevant, and that they could not comment on the effectiveness of the system because they were unable to interpret the results. In contrast the other non-Finnish/Swedish speaking participant appeared to have no difficulty retrieving and identifying relevant documents, and listed the most positive aspect of the system as being 'the translations from Swedish and Finnish'. In terms of precision and recall both participants were successful and could retrieve almost 1/4 of the total set of relevant documents; in addition they bookmarked as relevant almost 50% of the relevant documents displayed. This suggests that, despite user's impressions, the word-by-word translations were in fact accurate and substantial enough to support reasonable relevance judgements.

5. CLIR Usability: Some Key Points

The set of evaluations run during the Clarity project have been instrumental to raise our awareness on how CLIR should be designed to be usable. The need for "couching" the user's query towards terms that match the way the information is represented in the system has been recognised as a key point for a successful retrieval interaction (Belkin, 2000). This problem is accentuated in cross-language IR where the translation of the query adds a further layer of uncertainty. This section discusses elements that in our view are essential for an effective user-CLIR interaction.

The importance and the motivation for forcing user supervision over the query translation have been widely discussed in section 3. The interaction design proposed in section 4 mitigates the higher cognitive load required to the user as the two tasks of translation and search are kept together and are perceived by the users as a unit. By seeing the query translation users are more engaged with the search task and feel more in control. We regard this interaction proposal as fundamental though it uncovers potential weaknesses in the translation process that could undermine CLIR acceptability. Indeed we observed users frustrated from seeing incorrect or missing translations, and willing and capable of correcting the system. A usable CLIR should offer those skilled users the possibility of bypassing the translation step. In Clarity the '@' symbol is used to notify a translation bypass[5]: in the last evaluation a participant used "@research" while searching from Finnish to English as the translations proposed were not satisfying; in this way the user had forced the system to use a translation that was not in the dictionary but that was, in the user's opinion, more effective for retrieving relevant documents. Empowering the user over the CLIR system may mitigate the intrinsic problems pf query translation (Hull & Grefenstette, 1996).

The translation bypass feature is also valuable as English has infiltrated other languages and can therefore be used as pivot, e.g. 'computer' is used unchanged in other languages than English. This functionality directly derives from user requirements (Petrelli et al., 2002) as 'venture capital' (in English) was used to search Finnish databases. Again this feature may appeal to skilled users only, but this is a target new generation CLIR must consider, as discussed in the next section.

[5] The '@' symbol may not be the best choice as a user commented on its resemble with email and the Web.

A good dictionary feeds a good translation mechanism and is essential to offer users who do not know the target language a chance to retrieve relevant documents. It also means a more straightforward interaction as less query updating is needed and a more reliable result summary is displayed. Indeed the excellent dictionary used for translating English into Latvian (and back) allowed all users to assess the more diverse documents written in a problematical language. Other data has shown that CLIR is mature enough to support users with little or no knowledge of the target language in retrieving and, more importantly, identifying a significant proportion of relevant documents. Of course, as discussed in section 4, polyglots are better equipped and can fully exploit multilingual information access.

Last but not least a good query translation reinforces a sense of trustfulness essential when the goal of the user interaction is to retrieve information of potentially paramount importance, e.g. background for new business investments in foreign countries. A system that fails in translation and does not allow the user to fix it might be considered unreliable for effective use. Similarly a system that offers multilanguage IR should be consistent across its languages, for example translating geographical names for a language pair but not for another should be avoided.

A rich dictionary is a CLIR component worth investing on. However a good dictionary would not solve all problems related to crossing languages nor can we consider the problems related to CLIR solved just because a good dictionary is used. The experiments done seem to move the challenges for CLIR from linguistic aspects to cultural ones. An effective translation of proper names is the next frontier, particularly for languages with inflections or when a different alphabet is used and multiple transliterations[6] are possible. This would be a key to access news produced all around the world as dictionaries would not report names of current personalities.

Another challenge rarely considered but of potential impact for CLIR usability is the translation of phrases, particularly noun phrases. Indeed a phrase in the source language does not necessary match a permutation of the translated words in the target language. For example the English "green energy" better corresponds to the Italian "energia pulita" (literally "clean energy") then to the plain translation "energia verde". Moreover, even in the lucky chance of matching terms translation, prepositions might be introduced/excluded in the target language, for example the English "dry cleaning" translate in the Italian "pulizia *a* secco"; then a simple search for adjacent pairs would not be successful. All these aspects of daily use of language impact on the usability and must be considered by researchers if CLIR wants to be moved from labs into the real world.

6. Language, Culture, and Information Seeking: Some Reflections

Historically CLIR users have been considered people with limited language skills or people wishing to search multilingual databases (Oard 1997). Still today, users involved in experiments with CLIR systems may be required a poor or null knowledge of the target language (Dorr et al. 2003; Lopez-Ostenero et al. 2002). This seems bizarre considering that the majority of the world population is bilingual[7] (Baker, 2000) and that "approximately half of the world's on-line population speak a language other than English at home" (pg. 187, Baker, 2000). Polyglots are an enormous opportunity for multilanguage information access: they are well equipped for efficiently use it and they are potentially interested in multilingual content.

[6] Transliteration refers to phonetic translations across languages that use different writing systems.

[7] The term bilingualism encompasses many degrees of linguistic capabilities, from passive bilingualism (being able to understand and sometimes read in a second language without being able to speaking or writing in that second language), to biliteracy (being able to reading and writing in two languages), and biculturalism (besides the languages knowing the cultures of two different linguistic groups as well).

A questionnaire distributed during the initial Clarity user study showed as for specific professions language skills are required and it is not unusual to find people who are fluent in 4 or 5 languages. Those people use their skills for searching information, sometimes as often as everyday in the most diverse languages: for example one respondent to the questionnaire declared to search daily in Russian and English, once a week in German, French and Swahili[8], and occasionally in Farsi[9] and Chichewa[10]. Reasons for searching encompass collecting material for writing, finding information about people/companies/organizations, checking the correct spelling (places, people, organizations), checking fact and news.

Translators are another group of users potentially interested in CLIR. Their use of languages is extremely sophisticated as they have to render the language expressiveness (Eco, 2003). In our initial study translators were observed compiling own-dictionaries highly specific for the task in hand, e.g. Serbocroat[11]-English lists of religious and war terms. They also reported about the difficulties of translating idiomatic expressions as they often have completely different forms, e.g. the English "to beat about/around the bush" corresponds to the Italian "menare il can per l'aia" literally: walking the dog in the courtyard. In this context some form of collaboration between the user (who already manually constructs transfer dictionaries) and the CLIR system (that can exploit parallel corpora and show where the idioms were used) is worth exploring. It should be noted that good on-line dictionaries often include idiomatic expressions but those may be formatted differently from standard single words and are generally discarded in the translation process.

Other potential uses of professional multilingual information access include journalists checking daily how a piece of news is reported around the world and reports on business opportunities in foreign countries compiled by information specialists for investors. Both this scenarios derive from observed tasks: multilingual tasks exist but today those are carried out using monolingual tools.

Localization is a further example of potential for CLIR. Consider a user from a non-English speaking country typing a query to a Web search engine in English: that query could be automatically translated into the user's country language and used for searching the national domain for retrieving local instances of the required service. Again this application derives from a real need collected during the initial user study.

These many different examples suggest that more than "one size fits all" model, next generation of CLIR should target highly specialised applications that specifically address users needs and data characteristics. It is likely that promising uses (and market) have still to be discovered. As a further example consider multinational companies that produce goods and related documentation in many languages; some already use machine translation software[12] and are likely to have Intranet for better distributing that knowledge among the company premises. In this context the language is likely to be controlled and domain specific; as such a fuzzy translation for technical terms (Pirkola et al, 2003) can be the most appropriate tool and an effective CLIR system should be built around it.

A final reflection is on the social dimension and the global impact multilingual information access can have. With this respect, the Web has potential not exploited yet: it allows users from all around the world to retrieve information in the language where that is more available or reliable besides the

[8] Swahili is spoken on the east coast of Africa.

[9] Farsi (or Persian) is spoken in Iran and Afghanistan.

[10] Chichewa is widely spoken in south-central Africa.

[11] Serbocroat (or Serbo-Croat) was the official language of former Yugoslavia. After the Balkan conflict the two very closed languages Serbian and Croatian that formed it have been distinct.

[12] As claimed in the SYSTRAN case studies page http://translationsoftware4u.com/sys-testimonies.htm (accessed 10.3.2004).

language used as input. Consider for example medical information: it is unlikely that users would know medical terminology in other languages than their own, thought they can be able to read the retrieved documents.

Acknowledgements

Clarity is an EU 5th framework IST project (IST-2000-25310) and we gratefully acknowledge the support. Partners are: University of Sheffield (coordinator) (UK), University of Tampere (Finland), SICS – Swedish Institute for Computer Science (Sweden), Alma Media (Finland), BBC Monitoring (UK), and Tilde (Latvia). We are grateful to Jussi Kalgren and Premen Hansen for the help in collecting of part of the Swedish data, to George Demetriou, Patrick Herring, Heikki Keskustalo and Bemmu Sepponen for setting-up Clarity for the user tests. Finally we thank Alma Media and BBC Monitoring for the support during the final evaluation and all the people who participated in the three evaluations.

References

Baker, C. (2000) A Parents' and Teachers' Guide to Bilingualism. (2nd ed.), Multilingual Matters: Clevedon, UK.

Ballesteros, L., & Croft, W.B. (1998). Resolving ambiguity for cross-language retrieval. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (ACM SIGIR98). (pp. 64-71). Melbourne(Australia):ACM.

Belkin, N. (2000). Helping People Find What They Don't Know. Communication of the ACM. (43) 8, 58-61.

Borlund, P. (2000) Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, (53)3, 225-250.

Capstick, J., Erbach, G., Uszkoreit, H. (1998): Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents, Working notes of the AAAI Spring Symposium on Intelligent Text Summarisation. Stanfors(USA)

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), 275-289.

Chin, J. P., Diehl, V. A. & Norman, K. L. (1998). Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the CHI '88 Conference on Human Factors in Computing Systems, ACM Press, pp.213-218.

CLEF homepage http://clef.iei.pi.cnr.it:2002/ (accessed 1.3.2004).

Demetriou G., Keskustalo H., Sepponen B., Herring P., Franzén K., Karlgren J., Olsson F., Gaizauskas R. & Sanderson M. (submitted) A Web Services Architecture for Distributed Cross-Language Information Retrieval. Submitted to Journal of Natural Language Engineering.

Dorr B., He D., Luo J. & Oard D. (2003) iCLEF at Maryland: Translation Selection and Document Selection. In working notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway, 271-282.

Eco, U. (2003) Mouse or Rat? Translation as Negotiation. Weidenfeld & Nicholson

Gonzalo, J., Oard, D. (2002) The CLEF 2002 Interactive Track. In working notes for the CLEF 2001 Workshop, 19-20 September, Rome, Italy, 245-253.

Hull, D. & Grefenstette, G. (1996) Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. SIGIR 1996, Zurich, Switzerland, 49-57.

Levin S. & Petrelli D. (2003) Report on Effectiveness of User Feedback CLIR System. D6-2. Available at http://www.dcs.shef.ac.uk/research/groups/nlp/clarity/reports/d6-2.pdf (accessed 10.3.2004)

Levin S. & Petrelli D. (2004) Report on Effectiveness of Clarity System. D7-1. Available at http://www.dcs.shef.ac.uk/research/groups/nlp/clarity/reports/d7-1.pdf (accessed 10.3.2004)

Lopez-Ostenero F., Gonzalo J., Penas A. & Verdejo F. (2002) Interactive Cross-Language Searching: Phrases are better than terms for query formulation and refinement. In working notes for the CLEF 2001 Workshop, 19-20 September, Rome, Italy, 279-291.

McCarley, J.S., (1999) Should we Translate the Documents or the Queries in Cross-language Information Retrieval. In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics, 208 - 214.

NTCIR homepage http://research.nii.ac.jp/ntcir/index-en.html (accessed 1.3.2004)

Oard, D. (1997) Serving users in many languages cross-language information retrieval for digital libraries, D-Lib Magazine, December 1997.

Petrelli D., Hansen P., Beaulieu M, & Sanderson M. (2002). User Requirement Elicitation for Cross-language Information Retrieval. In The New Review of Information Behaviour Research, 3, 17-35.

Petrelli D., Beaulieu M, Sanderson M., Demetriou G., Herring, P & Hansen P. (in press) Observing Users - Designing Clarity: A Case Study on the User-Centred Design of a Cross-Language Information Retrieval System. Journal of the American Society for Information Science and Technology (JASIST) Special issue on "Document search interface design and intelligent access in large-scale collections"

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. & Jarvelin, K (2003) Fuzzy Translation of Cross-Lingual Spelling Variants. SIGIR03, 345-352.

Preece, J, Rogers, Y. & Sharp, H., (2002) Interaction design: Beyond human-computer interaction. Wiley.

Salton, G. (1973). Experiments in multi-lingual information retrieval, in Information Processing Letters, 2(1): 6-11.

TREC homepage http://trec.nist.gov/ (accessed 1.3.2004).

Xu, J., & Weischedel, R. (2000) TREC-9 Cross-lingual Retrieval at BBN. TREC 2000.