Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system

PETRELLI, Daniela <http://orcid.org/0000-0003-4103-3565>, BEAULIEU, Micheline, SANDERSON, Mark, DEMETRIOU, George, HERRING, Patrick and HANSEN, Preben

Observing Users - Designing Clarity
A Case Study on the User-Centred Design of a Cross-Language Information Retrieval System

Daniela Petrelli[*], Micheline Beaulieu[*], Mark Sanderson[*], George Demetriou[#], Patrick Herring[#]

[*]Department of Information Studies, [#]Department of Computer Science, University of Sheffield, Western Bank, Sheffield, UK

Preben Hansen

SICS – Swedish Institute of Computer Science, Stockholm, Sweden

Abstract

This paper presents a case study of the development of an interface to a novel and complex form of document 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. A study involving users (with such searching needs) from the start of the design process is described covering initial examination of user needs and tasks; preliminary design and testing of interface components; building, testing, and further refining an interface; before finally conducting usability tests of the system. Lessons are learned at every stage of the process leading to a much more informed view of how such an interface should be built.

1. Introduction

Cross language information retrieval (CLIR) is the retrieval of information written in one language based on a query expressed in another. For such a process to succeed, both translation and retrieval must be conducted in order for relevant items to be located. Although preliminarily explored by Salton (1973), CLIR was only researched in detail starting in the early 1990s with the EMIR project (Radwan, 1991). Research initially focussed on the matter of how to effectively translate between query (*source*) and collection (*target*) languages. A range of approaches were taken: using an existing machine translation system such as SYSTRAN (Gachot et al, 1995); extracting translations from bilingual dictionaries (Hull & Grefenstette, 1996); and training a translation system from existing translated corpora (Dumais et al, 1997).

Early research showed that CLIR was feasible but effectiveness was some way off that of equivalent so-called *monolingual retrieval* - classic IR where query and collection are written in a single language. With the establishment of a TREC track in CLIR, which led, within a few years, to a series of collaborative cross language efforts in Europe and Japan, retrieval of information written in a language different from the language of the query was researched widely. This resulted, by the late 1990s, in the effectiveness of CLIR for certain query-collection language pairs approaching that achievable with monolingual IR. Ballesteros and Croft (1998) were among the first to report CLIR effectiveness measured at over 90% of monolingual using a combination of automated query expansion (pre- and post-query translation) and a means of structuring queries so that translation alternatives of words were grouped. Other approaches that met or even went beyond monolingual retrieval were subsequently reported. Macarly (1999), for example, reported results suggesting that CLIR may be able to outperform monolingual retrieval alone. This was confirmed from the results in recent TREC efforts: for example, Xu & Weischedel (2000) reported exceeding by 18%, a simple monolingual retrieval using a complex query expanding CLIR method.

Cross language researchers have approached the problem of CLIR and, within a decade, succeeded in producing retrieval systems that, assuming translation resources of a certain quality exist, can retrieve as accurately as if query and collection were written in the same language.

1.1. Why is it not being used?

Despite the success of CL researchers in devising translation and retrieval methods that provide acceptable results, there are few, if any, public examples of CLIR systems. Web search engines, for example, although capable of searching in many languages and offering document translation facilities,

do not offer CLIR services. Quite why such an option is absent is not immediately obvious and is somewhat disappointing to CLIR researchers. While it is never easy to determine the precise reasons for search engine policy, we speculate it is due to the following reasons.

- Many searches on the web are for culturally specific information that are predominantly made by members of that culture. If a topic is specific to a culture, the majority of web pages relevant to that topic are likely to be written in the language of that culture. For the majority of users submitting such queries, cross language searching is unlikely to be beneficial and it is reasonable to assume that search engines are focussed on serving the majority interests of web users only.
- For queries that are not strongly attached to a particular culture, aspects of medicine, scientific research, configuration of computer operating systems, etc., it is conceivable that CLIR would be useful. However, it appears that while search engines are keen to increase their searchable collections, they rarely exploit means of altering queries, such as stemming or query expansion, to increase the number of pages matched. Some of the additional pages are likely to be relevant but most will not. As precision (the density of relevant documents returned) is generally more important to search engine users than recall (the total number of relevant returned), such *recall enhancing methods* are ignored by engine providers. Cross language retrieval can be viewed as such a recall-enhancing device: a form of query alteration to increase the number of matching documents. Consequently, it is unlikely to be adopted by IR systems that serve users with precision focussed information needs, as web searchers tend to be.

We contend that the domain of searching where CLIR may be of use is for queries that cross cultures and where recall is important. Tasks such as enterprise searching within multinational companies; patent searching; or serving the needs of the intelligence community may be examples.

1.2. Who wants CLIR?

A common thread in the CLIR academic research conducted is an almost exclusive use of a test collection methodology to evaluate retrieval approaches, with a collection of texts written in one language; queries written in another; and relevance judgements listing which collection documents are relevant to which queries. The effectiveness of CLIR systems was assessed using average precision. Test collection evaluation focuses on the initial retrieval and essentially ignores user interaction. Within CLIR, this is an important omission as it is unclear how users may interact with a CLIR system, what needs they may have, and more importantly, what sort of people wish to use a CLIR system.

There is not a great deal of literature on potential user groups of CLIR. Underpinning most of the research has been the assumption that users are searching for information written in languages that they do not read. It has been generally presumed users would have retrieved texts translated in some way. Consequently, testing of components that translate retrieved document lists is a common component of CLIR-usability testing (Gonzalo & Oard, 2002) and most research has focussed on small-scale usability testing of system components. A more substantial study was conducted for the MULINEX project (Capstick et al, 1998) where mock-ups of a CLIR system (German queries, English documents) were shown to 84 German university students whose preferences for means of result presentation were recorded in a questionnaire. One striking aspect of this study was the student's lack of interest in German translations of retrieved documents as they were all able to read them in their original (English) form.

Such an observation was not unique and in recent years, there has been a change in view on the type of user that might exploit CLIR: *polyglots*, people who speak more than one language, are a likely user group. If querying on some cross-cultural topic, they may find it annoying that they have to enter a separate version of the query into a search engine once for each language they know. Their needs as a user group are not well understood. This paper describes the user centred design of a CLIR system based on observations and interviews with such a group of bi- & tri-lingual users working in the media, e.g. journalists and broadcasters.

The work was conducted within a multi-site project called Clarity. Next section sets the context of user-centred system design, while the initial design and the field study are reported in sections 3 and 4 respectively. The data collected are analysed in section 5, and the effect on the system redesign are explained in section 6. Further refinement of the user interface and a set of user tests are discussed in sections 7 and 8. The final section, number 9, reports the conclusions.

2. User Studies: How and Why

To be effective, an information system has to be faithful to a real context and in keeping with the use the end-users will make of it. Designing with a user-centred approach requires that the user be involved during the whole design cycle (Norman & Draper 1986, Preece 1994). The process is iterative: after each important design phase, a user evaluation is performed and redesign follows. The process does not start with an implemented prototype, but with preliminary ideas of what the system should do, which are compared with what users are doing. Different techniques can be adopted at the requirements collection step (Schuler & Namioka 1993, Nielsen 1993, Hackos & Redish 1998); those used in the Clarity field study are discussed in Section 4 and are referred to throughout the whole design and implementation. When requirements are collected, there is the key step of transforming the findings into an interface and system.

Evaluations are used to test any kind of idea at any stage of the design-development process; for this reason they are called *formative evaluations*. They need to involve relevant users but can be performed using any prototype suitable for the purpose. Houde and Hill report (in Houde & Hill 1997) the use of a pizza box filled with bricks to simulate a laptop in size and weight. This form of low-*fidelity prototype* is useful to verify the effectiveness of ideas before any important choice about implementation is acted upon. In the case of Clarity, the low-fidelity prototypes were paper mock-ups, presented and discussed in Section 3.

Evaluations conducted with the prototypes are typically informal, with few participants, designed to find flaws in the system (e.g. looking for misleading interface elements or missing functionality). In our opinion, usability tests are appropriate when an exploration of alternative designs or conditions are the focus of the study. Even though Clarity is a research prototype and not a commercial product, we ran a considerable number of user tests in order to get a better understanding of the impact of the CLIR technology and to empirically select the best interface elements for such a special user interaction.

3. Creating a CLIR system: Early stages of Clarity

This section reports the user involvement at the beginning of the project when ideas were first generated.

3.1 Generating Scenarios

The first step was to determine the reality of users and uses in order to sketch a possible interface with a task in mind. An informal initial definition of users' needs was gathered via a discussion with end-users' representatives and lead to the writing of two *proposed scenarios* representing the designers' view of possible users, their tasks and their interaction with the future system.

Joannes Scenario

Joannes is a journalist. He is from Finland. He is fluent in English and French, but does not speak any Italian at all.	Any other relevant skills?
He has to write an article about the follow-up actions taken by the Italian government after the disorder in Genoa during last G8 meeting. He is interested in the political discussion that also followed those facts. Images are important too.	The task requires to find information on a single topic; only the most important ones matter.
He does not have a precise idea of what happened after. He heard rumours about public inquiry and parliament clarification done by the government after the Left asked for explanation.	
He sits at his computer and uses his usual browser to connect to CLARITY site. He is asked for login and password.	Clarity has to be multi-browser e.g.. Netscape and Explorer
He enters his domain, customised respect to his profile. Here the history of past searches is kept.	Is customisation important?
The screen for new search is displayed as default. Joannes types in few words: Genoa G8 disorder political discussion	Should Clarity translated proper names e.g. Genoa – Genova ?
Given his information need, his search will be more effective if Italian newspapers are searched. Nevertheless other international agencies may have reported about it already. So he sets the language to "search in" to Italian, but did not excluded other languages ("only" tick box left blank). ……	

Figure 1 An excerpt of a proposed scenario.

A scenario is a story describing a person with specific characteristics and motivations who performs a specific task by interacting with a specific system (Carroll, 1997). Scenarios were used in Clarity as a tool to stimulate discussion on what was feasible. Figure 1 shows an excerpt of such a scenario, on the left column the narration, on the right column the design questions. The right column was introduced to highlight the interface designer's questions on what Clarity should offer to users in term of system capabilities and interface features. Each question corresponds to a passage in the narration section.

Scenarios were used in Clarity to better understand the stages in the cross-language search task and as bases for the visualization exercise. Clarity scenarios were used to design a *proposed user interface*; mock-ups were drawn to address the different steps of an information searching behaviour.

3.2 Sketching the user interface

At this stage, designers examined past work. While little has been written about CLIR, a lot of empirically-based research in on general information retrieval task exists (Belkin et al. 1993, Brajnic et al 1996, Koenemann & Belkin 1996, Chen & Dumais 2000, Golovchinsky 1997, Cousin et al. 1997, Hearst 1999). On the basis of both the literature and the two proposed scenarios, different stages of the searching process of a user-CLIR interaction were identified and sketches of the stage were drawn.

1. **System setting-up**: users would work most of the time in the same conditions (same language in input, same language(s) in output, same collection etc.), thus a panel for setting up the system was considered essential even if rarely used;
2. **Query formulation and translation**: the user input should be supported by the system offering additional query terms and providing the user with query translation;
3. **Result overview**: given the potentially large retrieved set of documents and their heterogeneity, a graphical visualization of the whole set was considered desirable;

4. **Ranked list and single document inspection**: accessing a single document should be fast and direct from the ranked list itself;
5. **Multi-documents inspection**: document comparison was thought to be an important step when deciding documents relevance;
6. **Search history**: accumulating documents over search sessions was considered an important feature to be offered to users.

The six mock-ups generated are presented below. No attempt was made at this stage to relate the sketched *panes* to overall screen layouts. The hope was that participatory design sessions scheduled during the study would help in defining the final, global layout.

3.2.1 Query formulation and System set-up panes

Giving users control over a CLIR system means offering a mechanism to monitor and refine the query translation, as in both Arctos (Ogden & Davis 2000) and Mulinex (Capstick et al. 2000). Researchers assume users will enter queries in their language, which is translated by the system into the target language for retrieval (Oard 1997).

The initial design of the Clarity interface had a pane to allow typing in user control of the system's translation/expansion process. The *query formulation and translation pane* (fig. 2) was intended to serve this function. The first part of the pane displays a (non editable) summary of the current system settings. To edit those values, the user clicks on an arrow on the right upper corner; this action opens the system set-up pane (fig. 3). Below the setting summary, the query formulation section displays the translated query and other possible expansion terms suggested by the system. The user can include new terms in the final query as well as remove current translations using the two arrows displayed between the two lists. Terms displayed in the query translation window are in the target language and the back translation is included in brackets. Words with multiple meanings are expanded with a back-translated synonym so that the user can exclude those not related to the query. The example given in the figure is the term "disorder", which could mean "chaos" or "illness". When the user is satisfied with the translation, they can click a button to initiate a search.
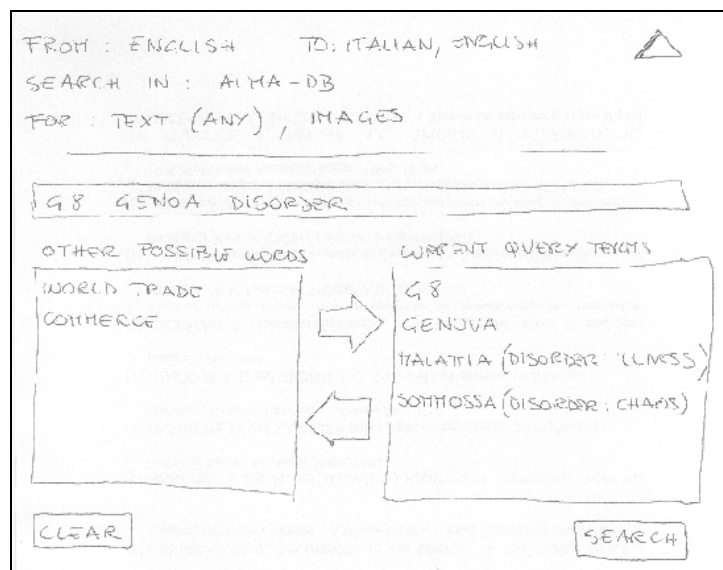

Figure 2. Mock-up of the query formulation and translation pane.

Figure 3. System set-up pane.

3.2.3 Result overview and ranked list

A *result overview pane* of the *retrieved set* can be represented graphically or content based (Leuski & Allen 2000, Chen & Dumais 2000). The ability to zoom in and out of a retrieved set through direct manipulation was deemed desirable by the designers. The cognitive complexity of searching cross-language let the designers hypothesise that at first, the predominant need would be to get an idea of the whole retrieved set. A graphical, highly interactive, representation of the ranked list was considered a starting point for the exploration of the retrieved heterogeneous set of documents. In the initial design, this visualization was the first part shown to users in the result pane (fig. 4). However, based on strong negative reactions from users, the overview was moved to a later stage of search: when a deeper investigation of the result maybe conducted.
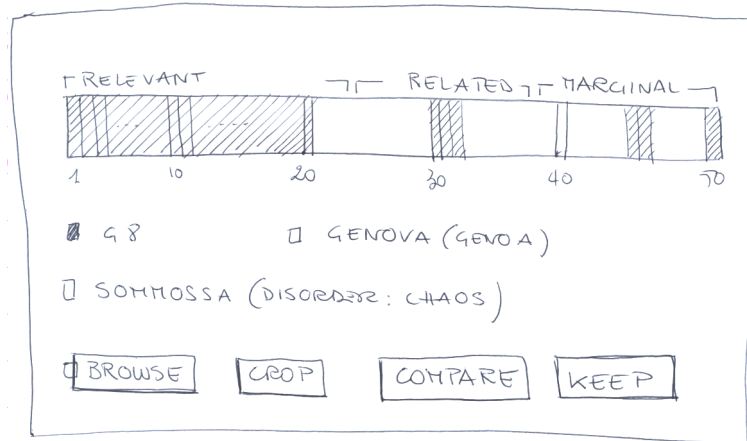

Figure 4. Result overview of the ranked list.

Clarity will also have a content-based visualization of the result based on the dynamic generation of a concept hierarchy (Sanderson & Croft 1999) built from retrieved document texts. The use of document derived concepts for organizing a ranked list has been shown to be effective: the work of Dumais et al (2001) shows the more information displayed, the better understanding the user gets. This work has influenced Clarity's ranked list display: the *result presentation pane* (fig.5). The layout provides a great deal of information on retrieved documents: title, document language, query terms in the document, significant key terms, both a document and user language summary, and a list of related documents. In this design, the set of retrieved documents can be sorted by document attributes, e.g. language, date, etc. Such a reorganization of the results is achieved by simply clicking on tabs

displayed at the top (figure 5 represents the list displayed respect to the "concepts" extracted from the documents).

Figure 5. Mock-up of the first sketch of the ranked list pane.

3.2.3 Search history

The final pane was the *working area*. This was envisaged as a part of the screen where the user could save accumulated documents and any other information considered useful. Few search systems provide support for keeping track of searches or saving queries and/or search results (Hearst 1999, Cousin et al. 1997). The users who participated in the study commented positively on idea. Despite the fact that more work is needed to clearly identify which parts of the search process needs to be retained and catered for by the interface, a minimal *working area* was tested with the users. It was implemented as a list of documents to keep throughout the next search.

4. The field study

Differences between users – their needs, tasks, colleagues, workplace, background - can make it difficult to design a single system valid for all. Equally, it is not realistic to design different systems to suit each user class. Therefore, some effort has to be spent in identifying the main user class(es) and to synthesise a set of features able to satisfy the broadest common uses. A field study was set-up with the to observe current practice on how real multi-language tasks are accomplished through monolingual tools and to imagine a multi-language system that would fully support multi-language information handling tasks.

4.1. Data Collection

A combination of data collection techniques was used in the field study, each capable of eliciting different types of information. The main technique used was *contextual enquiry*: observations of real users at work were undertaken simultaneously with interviews making it possible to focus on concrete

tasks and problems. The combined approach provided useful insight into the constraints of the environment as well as the dynamics of the social context where the action was taking place. Direct observations were conducted to get qualitative data for a limited number of subjects. To complement this technique, *questionnaires* were used to balance the limited view gathered by observing the users.

Users tried out a public CLIR system (ARCTOS[1]) and judged machine-translated web pages generated from Google. This collection of user feedback through *informal user evaluation* revealed further details about users' characteristics and the way they perform searching tasks.

In order to discuss design choices with end-users, a *participatory design* session was undertaken. Subjects were shown the six interface mock-ups and were invited to discuss and comment on the solutions presented even suggest alternative solutions.

4.2. User Participants

Two sites were involved: Alma Media, Helsinki, Finland and BBC Monitoring, Caversham, UK. Alma Media is a media company whose business is newspaper publishing, production and distribution of business information, television and radio broadcasting, and new media. BBC Monitoring supplies to English speaking customers news, information and comment gathered from news agencies and mass media worldwide. Across the two sites, 10 subjects participated in the study: 1 business analyst, 1 journalist, 3 librarians, and 5 translators. The strength of this type of study lay in the fusion of collected data, where even a single user counts in building a broad picture of requirements. On the one hand, it is the commonalities found across users and tasks, which form the basic skeleton of the interface design. On the other, the differences found between users are more likely to account for the provision of different options within the design to meet more diverse need (this point is discussed further below).

5. Data analysis and results

Data analysis involved making sense of all the observed situations and transforming findings to interface and system design choices. This section summarises the main findings and the implication for design (details in Petrelli et al. 2002).

5.1 Analysing the data

The main source of data for analysis was the video recordings complemented by the observer's notes and by the questionnaire. All user sessions were analysed to identify a number of points of interest as follows:

- **Goals**: final objective of the user's information seeking activity, e.g. write a report, translate a text.
- **Tasks**: addresses a set of coherent actions done for a purpose, e.g. find information about a person. Tasks could also be identifiable as sub-goals.
- **Acts**: the undertaking of a single atomic action, e.g. clicking an option to access a database, clicking on a retrieved document that seems promising.
- **Community context**: evidence of interaction between people, e.g. searching on behalf of a colleague.
- **Practices and procedures**: a common response to certain situations done despite its effectiveness, e.g. when results are not satisfactory changing the database instead of changing the query.
- **Design implications**: user suggested improvements related to existing design solutions, e.g. apply Boolean constraints over a retrieved set.
- **Opinions**: when the user expresses an opinion or preference, e.g. on the usefulness of a list of proper names extracted from the documents.

Examining the videos generated long lists of factual observations. The next step was to make sense of all the apparently unrelated data. As suggested by Hackos & Redish (1998) all the items in the lists were annotated and organized following the sequence of the observed users' tasks. This visual representation made it possible to identify commonalities across different users and tasks and relate them to the main system features, which supported them. For example, it was observed that all the users accessed different information sources in a single search session, although they may do so for different reasons. For example, the expert user exploited other sources to ensure a comprehensive

---

[1] http://messene.nmsu.edu/ursa/arctos/ accessed 10, Feb, 2002.

coverage, whereas a less experienced user accessed other sources as a result of a previously failed search, in either cases, such common behaviour must be supported.

5.2 User Classification

The first result to emerge from the data analysis was the creation of user classes: search experience, language knowledge, and final goal. As mentioned above, we met journalists, analysts, translators, and librarians. While the first three made a homogeneous group, librarians are different: they search on behalf of a customer who will make use of the retrieved information. Librarians might not know the language or the topic of the documents they are searching, and their final goal is to create an exhaustive list of documents to be delivered to the customer. By contrast journalists, analysts, and translators know the languages they are searching, and their final goal is to use the retrieved information. However their knowledge of search strategies and text collections can be limited, sometimes poor.

Thus language knowledge, search expertise, and final task (search-only or search-and-use) create different user classes with different user needs. Because of their differences with respect to the other groups, librarians were not considered primary users of Clarity (Hackos & Redish 1998). As a consequence, the interface design was not influenced by their specific needs. In the following only journalists and translators are considered[2].

5.3 User Requirements

A list of user requirements derived through comparisons and discussions was produced:

1. Users who know the language they are searching in (the majority) do so for other ultimate purposes (e.g. writing reports) and do not want (or know how) to control the translation or searching mechanisms;
2. Users want to search over multiple text collections and languages at the same time;
3. Users do not always use the most appropriate language they know for the task in hand;
4. English is used as a pivot to search other languages because of its relevance in technical jargon at international level; English can be used in combination with other languages;
5. Users would benefit from tools to sort results (e.g. by date, language, source etc.) and search within the retrieved set only;
6. Users often use compound names, proper names and phrases but have difficulties in generating synonyms term variants (e.g. venture capital, venture capitalist);
7. User-created dictionaries are a valuable support: the languages used by the same user for similar topics are generally the same.

The list affect different aspects of the system: a) the user interface (points 1, 3, 5, 7), b) generic mechanism of information retrieval (2, 6, 7), and c) specifics of the cross-language task (1, 3 and 4). Current cross language technology (Pirkola et al. 2001) satisfies points in b and c, the challenge is a: to provide what users want, making more intelligent use of the technology available based on a real context of use. For example, what seemed complicated to the observers, i.e. how to support a multi-language query, doesn't seem to be such a problem when the languages used are made explicit by the user, i.e. by ticking boxes.

6. Designing A New User Interface

Following the analysis of user information seeking, we recognised some of our initial assumptions were wrong. Sections 1 & 2 below report the redesign of query formulation and result presentation.

6.1 Query formulation

With the exception of an expert user, none of the subjects were interested in seeing how the system was translating the query. They were concerned with the search outcome alone, except when encountering difficulties in getting satisfactory results. Coupled with the problem that searching over many languages would have required a great deal of translation information to be displayed, led to the decision to remove the query formulation and translation pane.

As users easily switch from one language to another, having the languages fixed in separate windows (the set-up pane) was not a good solution. In the new version (upper part of fig. 6) the

---

[2] Business analysts were considered closer to journalists than to librarians even if their experience and knowledge as searchers was high.

separate summary for the multi-language search settings was removed and the remaining information embedded in the query formulation pane. In this way, users can move from one language to another, expand searches to further languages, or reduce the number of languages if the result of a query is unmanageable. This "language shift" was performed by all the observed users: a business analyst started a search using English, moved to Finnish and finally to Swedish, a journalist searched in a number of online newspapers using English, Spanish and German.

6.2 Result presentation

Users expect to be presented with a list of search results first. Any other form of graphical display is not what they expect. Consequently presenting an initial graphical overview of the results was set aside to be possibly used at a later stage in the search process.
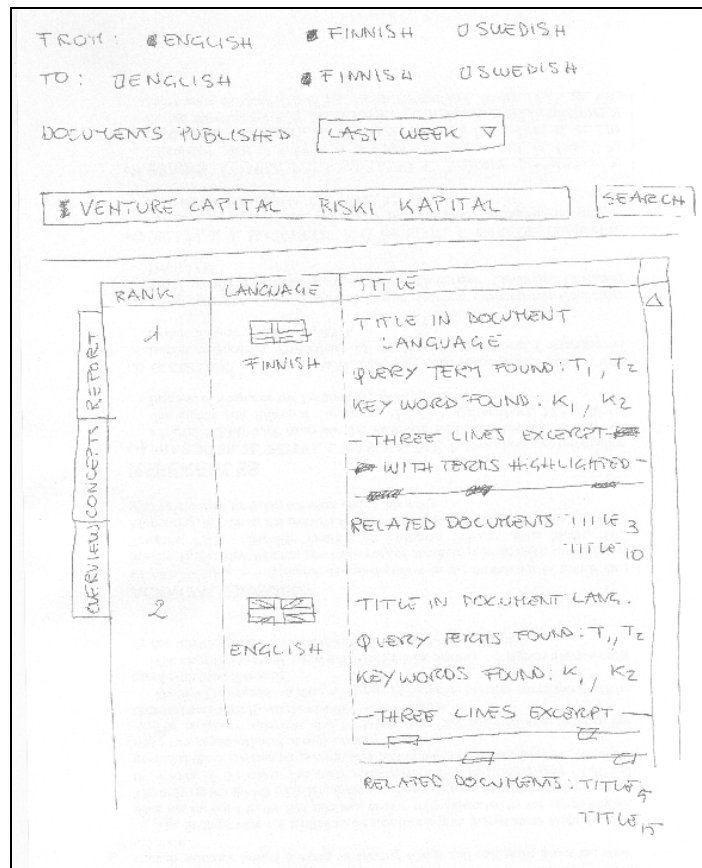

Figure 6. Mock-up of the revised interface of Clarity

In the revised version (fig. 6, lower part), the basic display was simplified. Because all the Clarity users are polyglots, there is no need for a summary of the document in the user language. Thus the possibility of giving the user the freedom to display more details or not (as in MULINEX (Capstick et al. 2000)) can be reconsidered. Features related to the content such as concepts or style are kept separate, accessed by vertical tabs. Result list sorting options on document attributes (date, language, etc) are selected by clicking on the head cell of the column (set horizontally on top of the ranked list).
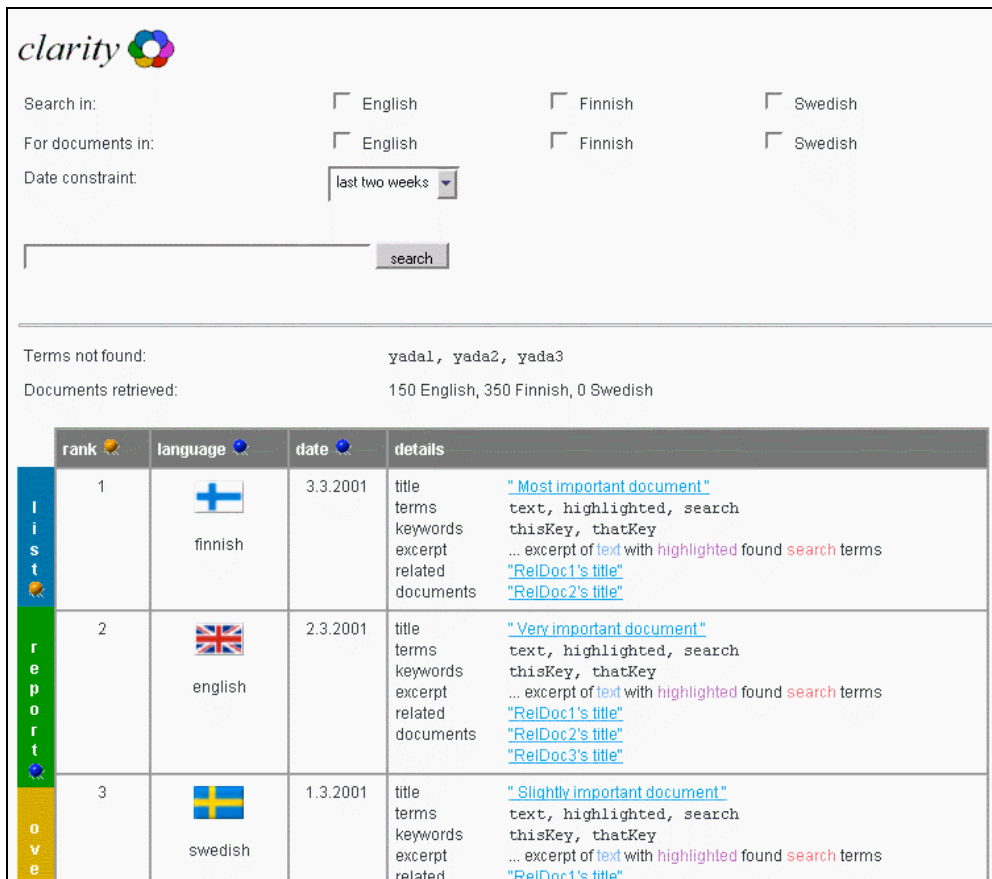
Fig. 8 The first implementation.

7. Interface refinements

Figure 8 shows the first implementation that reflects the design discussed above. Two HCI experts then tested this first prototype against a set of heuristics and following the scenarios (Nielsen 1993). The effect of seeing the interface on the screen with the right proportions revealed a few minor problems:

- *Tabs*: users expect tabs to be horizontal and at the top of a document (Krug 2000);
- *Alignment*: document titles are more important than their attributes (i.e. rank position, language), therefore, a left alignment of the title and a rearrangement of the key attributes was decided;
- *Similar documents list*: pointers to similar documents are space consuming and are not needed until a document is judged useful, thus those links were moved from the list, to the document display page.

The final layout, shown in Figure 9, was the one planned for the user evaluation, however some features were not provided by the Clarity search core system at that time of the first user interface implementation and had to be removed for the user test. Also, two graphic buttons were added: one to add a document to a "keep list" and one to see the query translation.
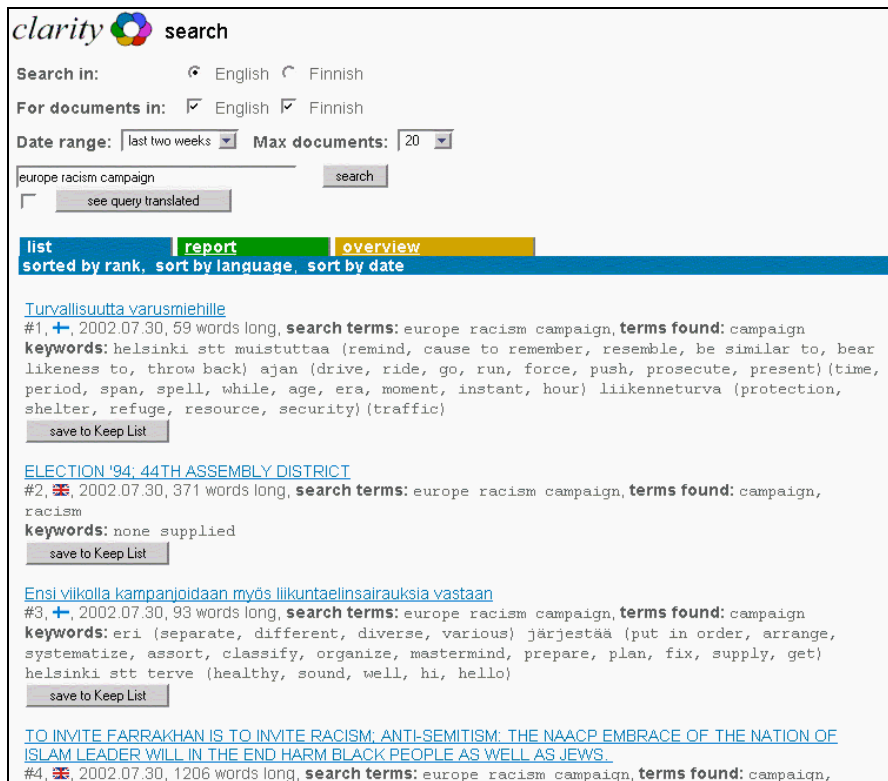
Fig. 9. The refined layout.

8. User tests

This section presents the background to and the user tests of the Clarity system.

8.1. Using empirical evidence to drive system design

As discussed above, the result of the field study was a new user interface that hid the query translation mechanism in favour of a simpler layout and interaction. Our solution is against the perspective adopted so far in CLIR systems and raises a set of fundamental questions:

- Is translation for cross language searching good enough so users do not need to supervise it?
- What is the implication of "hiding" or "showing" the query translation from a cognitive and problem-solving point of view?
- Does the knowledge of the target language impact on the needs and/or change the searching behaviour of users?

A set of user tests were undertaken from June to September 2002 with the aim of getting a better understanding of the cross-language searching task. At this stage, it was more important to get ideas for the design rather than to find a definitive answer on the best mode of interaction. Therefore several different and limited usability tests were preferred to a single, wider user experiment.

8.2. Conditions, Interfaces and Users

The choice of collapsing translation and search in a single step was well motivated by the result of the field study, but it has not been proved to be a real advantage. Therefore it was decided to compare the two conditions:

1) the user inputs the query, the query translation is shown by the system, the user verifies and/or modifies the query, and the system searches;
2) the user inputs the query, the system translates the query and searches without user intervention.

Two different input panes were used to test this condition: fig. 10a for condition 1 and 10b for condition 2. It should be noted that fig. 10a is an approximation of the query translation step showed in

fig. 2. It keeps the essence of the query translation check-and-revision step though, if revision is required, the user has to go back to the query input line and re-type the query.

The prime condition affecting only the input part of the interaction was correlated with two factors that effect the output: single/multi-language retrieval, and language competence. The first assessed if there was any difference between searching many languages at a time versus searching only one; the second considered the effect of user language skill (poly or monoglot). To support the monolingual users, the basic interface was changed. Document titles in Finnish were supplemented with word-by-word translations in English. The same mechanism was used to translate the keyword list.

The interfaces used for testing the different output conditions were: multilingual output for polyglots, (bottom fig. 9); single language for monolingual users, (bottom fig.10b); and multilingual output for monolingual participants, a combination of the output in fig. 10b for Finnish and fig. 9 for English.
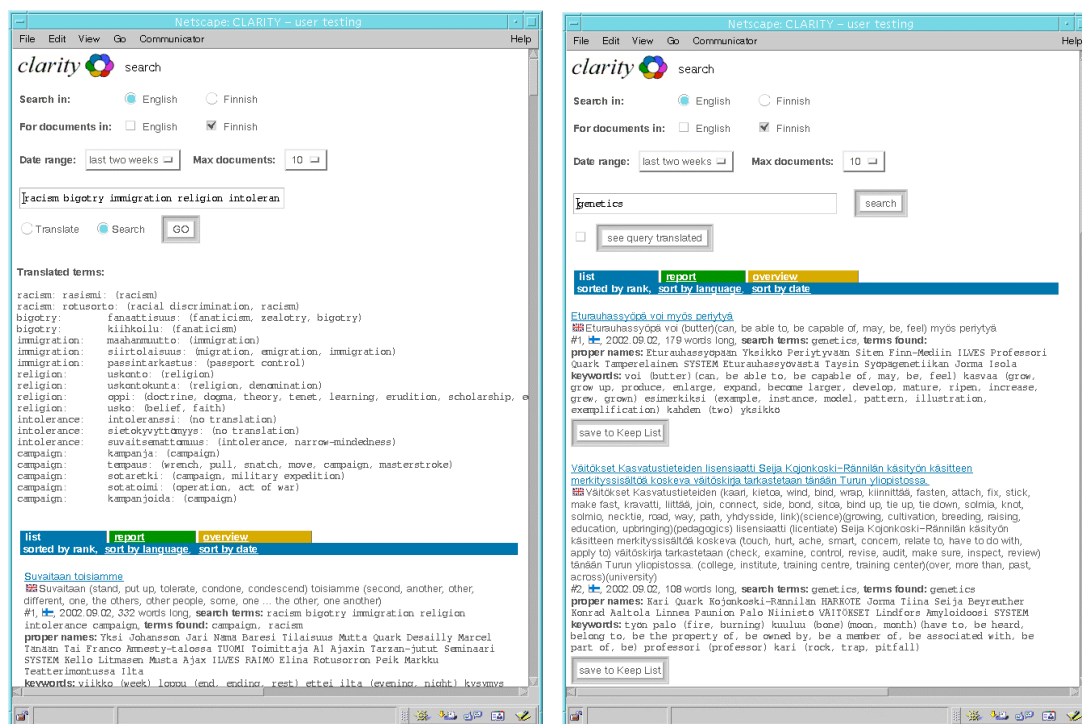

Figure 10. Layouts in single language: a) force check translation; b) hidden translation.

Ten people participated in the tests: six were monolingual (English) who tested the single language output; two were polyglots (Finnish, first language, English second) testing the multi-language output, and two were monolingual users (English) testing the multilingual output.

On arrival, participants received a description of the experiment and its purpose. An initial questionnaire was filled in, which was designed to collect user expertise with computers and searching and to ascertain their attitude to IR. Next, two training searches on each system were performed. Participants were provided with a scenario that described the task in hand (Borlund & Ingwersen 2000). During the tests, each participant used both input interfaces while testing only one output condition. Participants were required to select relevant documents by saving them in the keep list. Questionnaires were collected after each task, after a session with a system and at the end of the test. A log recorded the queries submitted, the retrieved set, and the relevance judgement given by users. Whenever possible, participants were videotaped or observed by the experimenters in order to collect qualitative information about the interaction. The procedure mirrored as closely as possible Gonzalo & Oard's iCLEF experiments (2002).

8.3. Result discussion

In the testing, the whole Clarity prototype was used. The system was physically distributed with a co-ordinating server and interface at the University of Sheffield and cross-language services (query

translation and retrieval) at the University of Tampere. SOAP[3] was used to facilitate communication. A technical problem hampered some data collection, which had the effect of pushing participants to try as many search strategies as they could.

The questionnaires revealed that almost all subjects preferred the interface that hides the translation even if the difference between the two was rated as minimal.

8.3.1. Initiating a query

Query formulation is always affected by a user's searching strategy. In cross-language, the cultural background can bias the search. For example a participant was observed typing "bobby sand" while searching for "hunger strikes". While the searcher presumably knew that Bobby Sands was a hunger striker, the system searched for Finish translations of "policeman" and "beach". Such searching by example may commonly fail in CLIR because it is a cross-cultural task.

Another result was that users expect the system to support a more sophisticated input than a simple list of words, which was all the system supported. Participants tried to specify phrases using inverted commas, mandatory words using '+', and synonym lists, e.g. "DNA", "D N A", "D. N. A.".

8.3.2. The effect of seeing the query translation

Almost all the observed people were puzzled when the system showed the query translation. They expected to see the result of the search and waited (for a while) for it to appear. When the translation was shown, all modified the query focussing it in some manner thereby preventing the query in its initial form from being run. Such a strategy might prevent the retrieving of relevant documents only matching the generic terms.

A similar phenomenon occurred when the user who was interacting with the interface (shown in fig. 10b) explicitly asked to see the query translation (by selecting the "see query translated" button). In this condition, the translation is shown but the search is automatically done; when the participants saw multi-sense words they wanted to stop the system in order to change the query, an action that was not allowed by the interface. However, when the result was presented, participants paid attention to the documents so they did not go back to reformulate the query but explored the retrieved set. A new interface design worth testing would be to hide the translation completely and offer it as additional pane available for inspection when the retrieval has already taken place.

A new layout for the input condition that requires the user checking the query translation (input condition 1) was also suggested by the tests result. When facing ambiguous terms, participants tried to disambiguate the query by using the Finnish word corresponding to the sense of interest. This made the system perform worse because Finnish words were mixed with English words when only one language can be used as input. Users became frustrated, which could have been avoided by a layout that allowed user selection of the right sense, as suggested by a couple of users.

Almost all users complained about the large number of translations of certain query words, in particular those resulting from figurative interpretations of words. For example when typing "green" none of the participants expected to retrieve documents about golf. Limiting the translation of generic and polysemous words to only the most common terms will on the one hand present the query translation in a more compact way and on the other hand make the searching more effective since the distortion introduced by uncommon senses will be automatically removed.

8.3.3 Looking at the search result

In the multi-language condition, participants complained about the order of the retrieved list, which was a simple interleaving of English and Finnish documents. The possibility of resorting the results by language was offered by the interface, but not used. Even if a interface redesign is needed to make the sorting options clearer, the default for multilingual output will be in future by language.

While viewing multilingual documents, knowing the target language affects users behaviour. Finnish speakers only read the title of the Finnish documents and did not open them to look in depth; conversely the same people opened the English documents to get confirmation of their first impression. This behaviour might have been due to the relative ranking of the two sets retrieved from the two collections, but it might also show a higher degree of uncertainty when judging documents in another language. English monolingual participants under the same condition showed the opposite behaviour. Most of the time they did not open the Finnish documents but based their judgement on the summary

---

[3] http://www.w3.org/TR/SOAP/ accessed 1, Oct, 2002.

provided, while often they opened the English ones. We think that the influence of the language on user behaviour when searching cross-language is not clear and needs further investigation.

Different opinions have been collected with respect to the information on each document in the ranked list: a word count was considered by two subjects as irrelevant while another used the attribute for deciding if the document was long enough to be of interest. The keyword list in particular seems to be a controversial point. Participants who speak Finnish considered it excessive; by contrast, non-speakers used the translated words to decide about the document interest. In this context, customisation may be an advantage.

Clarity works in a standard way, in that the same document can be retrieved for different searches even if it has already been retrieved. Participants preferred not to see again documents they had already viewed. This feature was particularly annoying due to the test system's slow search. A mechanism for either indicating, partitioning, or removing previously viewed documents will be considered.

9. Conclusions

The user tests have provided a better understanding of the cross-language task. The intuition derived from the field study that the query translation should not be considered as a separate task seems to have been confirmed. Seeing the translation affects searching behaviour and also potentially affects CLIR system performance. We consider this result an important finding for cross-language research. If CLIR engines can perform as efficiently as the monolingual-retrieval, query translation and search can be considered a single step unsupervised by the user. However, this might be true only with systems performing exceptionally well. It might be that by seeing the query translation the user is encouraged to revise and rethink the query much more than they would do if the translation were hidden. A modified query might be more efficient in retrieving the relevant documents than the initial one. If the initial query was submitted and the user is happy with the documents retrieved, it is unlikely that they will change the query to make it more effective. It might also be the case, as observed in the field study, that the user does not know how (or why) to change the query. Taking all this into consideration it might be that showing the translation forces the user to apply more efficient strategies than they would do so if they were left alone. A recent experiment by He et al. (2002) indicates this contradiction of users being more effective in retrieving documents when the translation is shown but liking it far less than the one that hides the translation. However, He's experiment considers a CLIR system with a single target language. Searching many languages means showing many query translations simultaneously and requires a more complex layout.

On the bases of results to date the next step is to undertake a broader evaluation to explore the tension between cognitive mechanisms and system efficiency. In particular, there is the need to address the three evaluation criteria of usability (namely efficiency, effectiveness and user satisfaction) in CLIR and if this changes if the target language is one or several.

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

Beaulieu, M. and Jones, S. (1998) Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers*, 10(3), 237-248.

Belkin, N. J., Marchetti P. G. and Cool., C. (1993) BRAQUE: Design of an interface to support user interaction in Information Retrieval. Information Processing & Management, 29(3), 325-344.

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

Brajnic, G., Mizzaro, S and Tasso, C. (1996) Evaluating user interfaces to information retrieval systems. A case study on user support. In: Frei, H-P., Harman, D., Schäuble, P., and Wilkinson, R.

(eds.). *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96).* Zurich, Switzerland. August 18-22, pp. 128-136.

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, in *AAAI Spring Symposium on Intelligent Text Summarisation*

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.

Carroll, J. M. (1997) Scenario-based Design, In M. Helander, T. Landauer, and P. Prabhu eds. Handbook of Human-Computer Interaction, second edition. Elsevier, 384-406

Chen, H. and Dumais, S. (2000) Bringing Order in the Web: Automatically Categorizing Search Results. Proceeding of CHI2000, 145-152.

Cousin, S. B., Paepcke, A., Winograd, T., Bier, E. A. and Pier, K. (1997) The Digital Library Integrated Task Environment (DLITE). In: Proceedings of DL 97, Philadelphia: PA, 142-151.

Dumais S., Letsche, T., Littman, M., Landauer, T. (1997) Automatic cross-language retrieval using latent semantic indexing. In AAAI Symposium on CrossLanguage Text and Speech Retrieval. American Association for Artificial Intelligence.

Dumais, S., Cutrell, E. & Chen, H. Optimizing Search by Showing Results in Context, Proceedings of CHI International Conference on Human Factors in Computing Systems (ACM SIGCHI '01) (pp. 277-284).

Dumas, J. S. and Redish, J. C. (1999) A Practical Guide to Usability Testing, Intellect.

Gachot, D. A., Lange, A., Yang, J. (1998) The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval. In G. Grefenstette, editor, Cross-Language Information Retrieval. pp.105-118.

Golovchinsky, G. (1997) Queries? Link? Is there a difference? Proceedings of CHI97, Atlanta: GE, 407-414.

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

Hackos, J. T. and Redish, J. C., (1998) User and task analysis for interface design.Wiley.

He D., Wang J. Oard D. & Nossal M. (2002) Comparing user-assisted and automatic query translation. Working notes for the CLEF 2002 Workshop, 267-278.

Hearst, M. A. (1999) User Interfaces and Visualization, Chapter 10 in Baeza-Yates R. And Ribeiro-Neto B. 'Modern Information Retrieval, Addison-Wesley, 1999. (available at http://www.sims.berkeley.edu/~hearst/irbook/chapters/chap10.html accessed 1.2.2002)

Hull, D.A., Grefenstette, G. (1996): Querying across languages: a dictionary-based approach to multilingual information retrieval, in *Proceedings of ACM SIGIR Conference*, 19: 49-57

Houde, S and Hill, C. (1997) What do prototypes prototype? In M. Helander, T. Landauer, and P. Prabhu eds. Handbook of Human-Computer Interaction, second edition. Elsevier, 367-382.

Leuski, Anton & Allen, James (2000) Lighthouse: Showing the Way to Relevant Information. IEEE Symposium on Information Visualization 2000 (INFOVIS 2000), Salt Lake City: Utah, October 9-10 2000, 125-130.

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

Koenemann, J., & Belkin, N. J. (1996) A case for interaction: A study of interactive information retrieval behavior and effectiveness, Proceedings of CHI96, 205-212

Krug, S. (2000) Don't Make Me Think – A common sense approach to web usability. Indianapolis: New Riders.

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.

Nielsen, J. (1993) Usability Engineering. Academic Press.

Norman, D. (1986) Cognitive engineering. In: D. Norman and Draper, S., eds. *User centered system design. New perspectives on human-computer interaction*. Hillsdale, N. J.: Lawrence Erlbaum Ass., 31-61.

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

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

Petrelli D., Hansen P., Beaulieu M, Sanderson M. (2002) User Requirement Elicitation for Cross-language Information Retrieval. 4th International Conference on Information Seeking in Context - ISIC 2002.

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

Preece J. (1994) Human-Computer Interaction. Addison-Wesley

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

Radwan, K., Foussier, F., Fluhr, C. (1991): Multilingual access to textual databases, in *Proceedings of RIAO 91, Intelligent Text and Image Handling*: 475-489

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

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

Schuler, D. and Namioka (eds.) (1993) Participatory design: principles and practices, Hillside: Lawrance Erlbaum Ass., 1993.

Xu, J., Weischedel, R.: TREC-9 Cross-lingual Retrieval at BBN. TREC 2000