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

Daniela Petrelli[*], Preben Hansen[§], Micheline Beaulieu[*], Mark Sanderson[*], George Demetriou[#], Patrick Herring[#]
[*]Department of Information Studies
[#]Department of Computer Science
University of Sheffield, Western Bank, Sheffield, UK

[§]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, e.g. typing a query in English to retrieve documents written in Finnish. For such a process to succeed, both translation and retrieval must be conducted in order to locate relevant items. Although preliminarily explored by Salton (1973), CLIR was only researched in detail in the early 1990s (Radwan et al., 1991). Research initially focussed on how to effectively translate from a query (*source language*) to a collection (*target language*). A range of approaches was adopted including: using an existing machine translation system such as SYSTRAN (Gachot et al, 1998); 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 *monolingual retrieval,* i.e. classic IR where query and collection are in a single language. With the establishment of a TREC track in CLIR, which led 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 more widely. This resulted 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 retrieval by 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. McCarley (1999), for example, reported results suggesting that CLIR may be able to outperform monolingual retrieval alone. This was confirmed by the results in recent TREC efforts: for example, Xu & Weischedel (2000) reported improving a simple monolingual retrieval by 18% by using a complex query expansion CLIR method.

While early research focused on retrieval functionality and effectiveness, no attention was paid to potential utility of CLIR and users were rarely involved in evaluation. Moreover the main assumption underpinning the research was that users of CLIR do not have any knowledge of the target language. Hence this would require a limited translation at the retrieval stage in order to *detect* relevant documents in the target language(s), which could then be fully translated by a human. For example, the retrieved document list could be translated but not the full documents and users could be involved only in assessing the retrieved document list (Oard & Gonzalo, 2001). However when a more holistic user study was conducted for the MULINEX project (Capstick et al, 1998), subjects demonstrated a lack of

interest in the German translations of retrieved documents as they were all able to read the original (English) form. This illustrated the use of CLIR by a different user group, i.e. *polyglots* people who speak more than one language. Polyglots do not need document translation but can benefit in querying cross-cultural topics since it is likely that such a group would find it annoying to enter a separate version of the query for each of the languages concerned.

It would thus appear that different user groups demonstrate different behaviours and this has implication for the design of effective CLIR systems. This paper describes a 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 approach is discussed in the next section together with an overview of the paper.

2. User Centred Design

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, Preece et al. 2002). 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 stage (Schuler & Namioka 1993, Nielsen 1993, Hackos & Redish 1998). The key step of transforming the requirements into the design of interface and system features is mainly a matter of creativity. However some guidelines can help in directing the effort (Preece et. al. 2002). Evaluations are used to test any kind of idea at any stage of the design-development process. They need to involve relevant users but do not necessarily require a full working prototype. Such evaluations are typically informal and formative, with few participants, run in the lab, and designed to find flaws in the system (e.g. looking for misleading interface elements or missing functionality).

The work discussed in this paper used a user-centred design approach to lead the development of a CLIR system, called Clarity, toward its first stable prototype. The process included four phases as follows:

1. **Scenarios and Preliminary Design**: an informal initial definition of users' needs started the process; scenarios were written, representing the designers' view of possible users, tasks and interaction, and were the basis for sketching mock-ups of a proposed user interface. This first step is discussed in section 3.
2. **Requirements Specification**: direct observation of users at work were conducted to get as much detail as possible on real users performing real tasks in real environments; the mock-ups were submitted to users' judgement in participatory design sessions; other additional techniques, i.e. interviews, informal user evaluations and questionnaires, were used to complement the main data collection. Section 4 reports about the study while section 5 gives details on the analysis.
3. **Design**: the integration and combination of the results of the previous phase led to a revision of the first mock-ups and a new solution was reached based on the knowledge acquired. This process is described in section 6, while a refinement step is briefly commented on in section 7.
4. **Formative Evaluation and Redesign**: usability tests were used to explore alternative solutions and to get a better understanding of the impact of the CLIR technology. Results were extremely interesting, even if not definitive, and redirected both research and design, as discussed in section 8.

3. Creating a CLIR System: Early Stages of Clarity

The aim of the Clarity[1] project was to create a CLIR system for rare languages, i.e. those with few electronic resources. The prototype discussed here used English as the source language and Finnish as the target language.

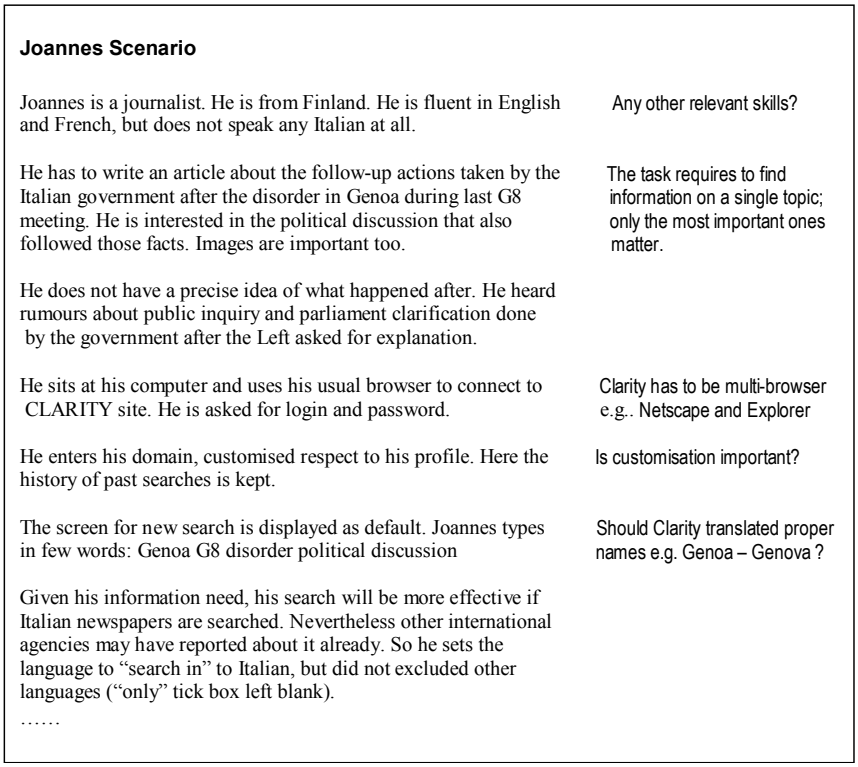
[1] Clarity website is http://clarity.shef.ac.uk/

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 who are part of the Clarity consortium, namely BBC Monitoring (UK) and Alma Media (Finland). BBC Monitoring supplies news, information, and comment collected from news agencies and mass media worldwide to English speaking customers; Alma Media is a media company whose business is newspaper publishing, production and distribution of business information, television and radio broadcasting, and new media.

The user groups initially considered were journalists and business analysts, with the intention of expanding to other categories (i.e. translators and librarians, see 4.2). Other professions are also potential users of CLIR (e.g. intelligence and patent agents) but these were not included due to logistic and content restrictions (e.g. the lack of language pairs relevant for intelligence and patent tasks).

The initial discussion led to the writing of two *scenarios* representing the designers' view of possible users, their tasks and their interaction with the future system (Carroll, 1997, Rosson & Carroll, 2002). In the first one, a journalist had to search an event that occurred in Italy without knowing Italian (see fig. 1); he relies on his knowledge of other languages and on the international relevance of the event. The second scenario depicts a business analyst who monitors fortnightly the news on a set of predefined Spanish companies; she knows the topic and the language well and the best search strategy, i.e. retrieve documents generated in the last two weeks.



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 scenario.

Scenarios were used in Clarity as a tool to stimulate discussion on what was feasible. Figure 1 shows an excerpt with the narration in the left column and the design questions in the right column. The right column was introduced to highlight the questions on what the system should offer to users in terms of functionality and interface features. Each question corresponds to a passage in the narration section.

Besides supporting technical discussion among people with different backgrounds, scenarios were used to better understand the stages in the cross-language search task and as a basis for the visualization exercise described in the next section.

3.2 Sketching the User Interface

At this stage, designers examined past work. While little has been written about CLIR interaction, a lot of empirically based research on general information retrieval task exists (Belkin et al. 1993, Brajnic et al. 1996, Koenemann & Belkin 1996, Beaulieu 1997, Cousin et al. 1997, Golovchinsky 1997, Beaulieu & Jones 1998, Hearst 1999, Chen & Dumais 2000). 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 were drawn. The prototypical process imagined, extensively revised later, was as follow:

1. **System setting-up**: users would work most of the time in the same conditions (same source language, same target language(s), same collection(s) 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 document relevance.
6. **Working area**: accumulating documents over search sessions was considered an important feature to be offered to users.

During a brainstorming session three interface designers went through the six steps and generated six mock-ups (see below). Some attempt was made to define an integrated graphical interface (as those proposed by Cousin et al. 1997 and Hendry & Harper 1997). However no decision was taken since a full understanding of the user-CLIR interaction was not yet clear. Thus sketches were kept distinct and the global layout left for further discussion at a later stage.

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 the Keizai (Ogden & Davis 2000) and Mulinex (Capstick et al. 2000) systems. Researchers assume users will enter queries in the source language, which is translated by the system into the target language for retrieval (Oard 1997).

The initial design of the Clarity interface followed this well traced path and allowed users to fully control the query translation process. Query expansion has been shown to be successful in CLIR (Ballesteros & Croft 1998) thus that feature was included to better support the *query formulation and translation*. The sketched pane (fig. 2) displays a (non editable) summary of the current system settings. Below it, the query formulation section displays the translated query (list on the right) and other possible expansion terms suggested by the system (list on the left). The user can include new terms in the final query as well as remove current translations by using the two arrows displayed between the two lists. Query terms translated into the target language also include the source term in brackets. Words with multiple meanings are expanded with synonyms so that the user can exclude those not related to the query. In figure 2 the ambiguous term “disorder” is expanded with “chaos” and “illness”.

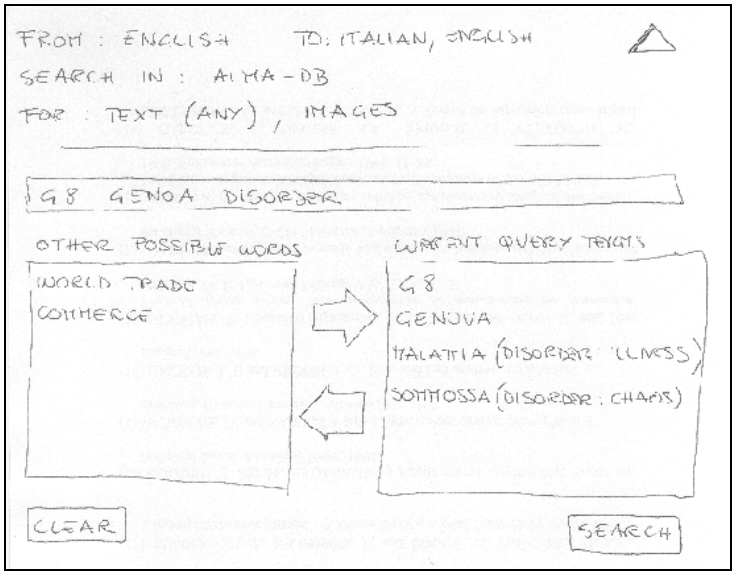


Figure 2. Mock-up of the query formulation and translation pane.

Clicking on the up arrow on the top opens the *system set-up* pane (fig. 3) and offers options to change the general conditions of the search, e.g. source and target languages, text collections. Filters on document span-date (from-to) and media types (text, image, voice, video) were also included since those attributes were considered important features to focus the search from the very beginning.

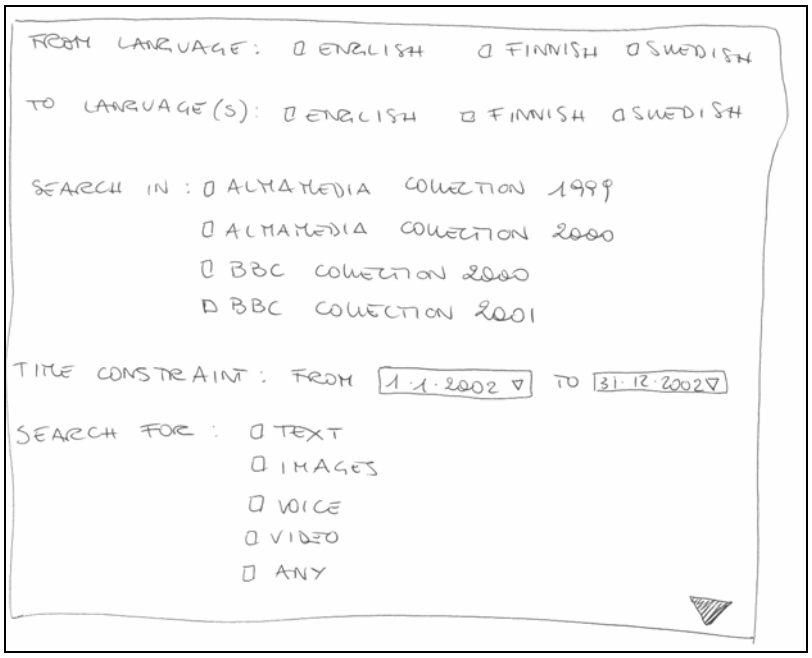


Figure 3. System set-up pane.

3.2.3 Result Overview and Ranked List

A *result overview* pane of the retrieved set can be represented graphically or with text (Leuski & Allen 2000, Chen & Dumais 2000). 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 complemented with the possibility of zooming in and out through direct manipulation. Thus 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 result shown (fig. 4). However, based on strong negative reactions from users the idea was temporary set aside (see section 6.2 for the discussion).

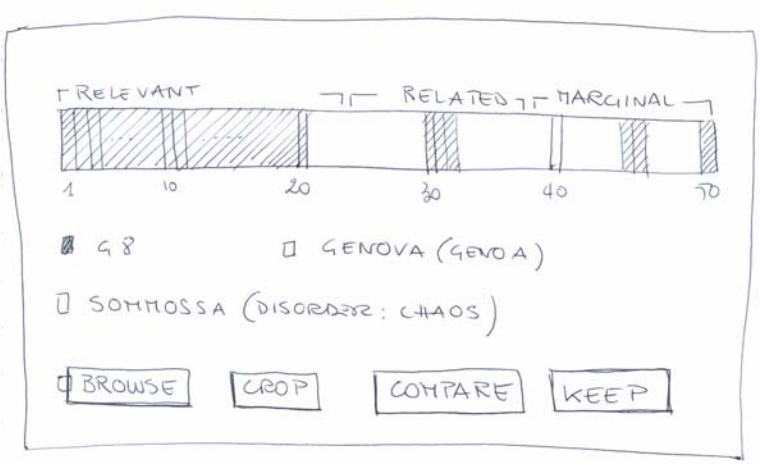


Figure 4. Result overview of the retrieved set.

Clarity will also have a content-based visualization of the result based on the dynamic generation of a concept hierarchy built from retrieved document texts (Sanderson & Croft, 1999). 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 that the display of more information leads to a better understanding by the user. This work has influenced Clarity's initial *ranked list* display (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 target and source language summary, and a list of links to related documents. In this design, the set of retrieved documents can be sorted by attributes, e.g. language, date, etc. Such a reorganization of the results is achieved by simply clicking on tabs displayed across on the top (figure 5 represents the list displayed with respect to the "concept" extracted from the documents).

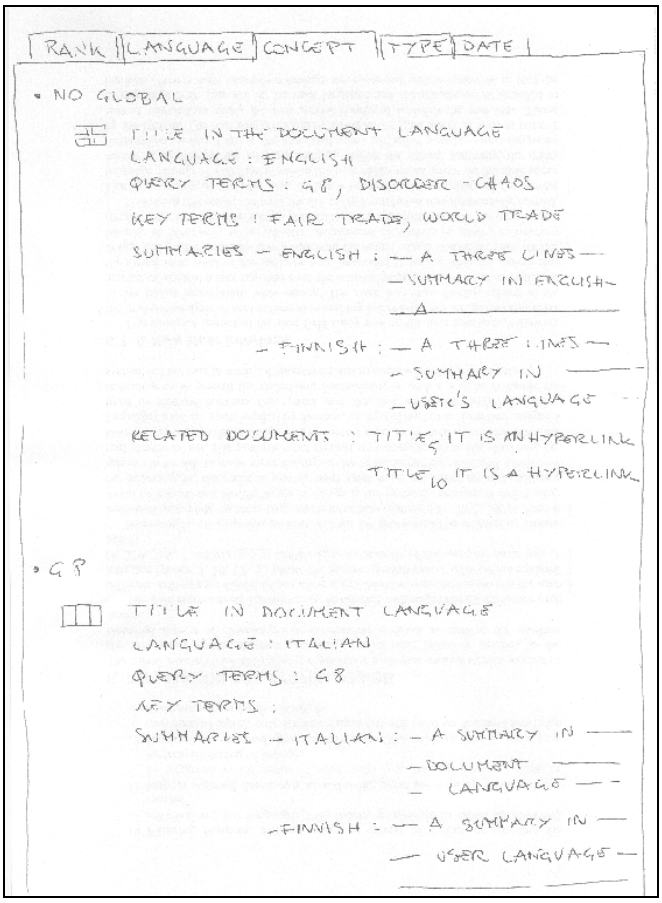


Figure 5. Mock-up of the first sketch of the ranked list pane.

In addition to a simple document display, a document comparison window was considered (Ogden & Davis 2000, Golovchinsky 1997). A pane was designed to display a pair or four documents simultaneously and to support browsing through the "related documents" list. Despite user interest, this feature was temporary abandoned due to the current limitation of the searching system in not providing document similarity searching.

3.2.4 Search History

A *working area* was envisaged as a place where the user could accumulate 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 the 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, thus called "keep list".

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 to observe current practice on how real cross-language tasks are accomplished through monolingual tools.

4.1. Data Collection

A combination of data collection techniques was used in the field study, each capable of eliciting different types of information. All sessions were videotaped.

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. Participants carried out their own tasks at their desk using their own tools and strategies; observers sat alongside them and questioned them on work organization, cooperation with colleagues, tools used, and strategies adopted. The combined approach observation-interview provided useful insight into a range of aspects of cross-language information seeking, including the environment and the dynamics of the social context where the action was taking place.

Contextual enquiries gave qualitative data for a limited number of subjects; this was complemented by *questionnaires* used to balance the limited view, e.g. the need for multi-language search. The questionnaire collected data on linguistic competence and use, search experience in general and searching cross language in particular. A final session aimed at measuring the actual search competence, e.g. knowing what a ranked list is.

Participants were also asked to try out a public CLIR system (Arctos[2]) and to judge machine-translated web pages generated from Google. This collection of user feedback through *informal user evaluation* revealed further details about users' characteristics including: the way they perform searching tasks, their searching competence and strategies adopted.

In addition 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 and also suggest alternatives.

4.2. User Participants

Two Clarity partners, Alma Media and BBC Monitoring, kindly supported the study. In total 10 subjects participated: 1 business analyst, 1 journalist, 3 librarians, and 5 translators. The strength of this type of study lies in the accumulation of 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 forms the basic skeleton of the interface design. On the other hand, the differences found between users are more likely to account for the provision of different options within the design to meet a diversity of needs.

---

[2] Arctos was accessed from the Ursa Project page at http://crl.nmsu.edu/research/Projects/tipster/ursa/ last 10th Feb 2002. Screenshots of a typical search are available at http://crl.nmsu.edu/~ogden/I-clir/clir-interactive/arctos/page1.html

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 the design (see 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 undertaken 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 undertaken irrespective of 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. adding Boolean operators
- **Opinions**: user expresses an opinion or preference that can lead to new design solutions, 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. The expert user exploited other sources to ensure a comprehensive coverage, whereas a less experienced user accessed other sources as a result of a previously failed search, in either case, such common behaviour must be supported.

5.2 User Classification

We met journalists, analysts, translators, and librarians and discovered they differ in search experience, language knowledge, and final goal. While the first three made a homogeneous group, librarians are different in that they search on behalf of a customer who will make use of the retrieved information. Librarians might not know the language nor the topic they are searching for, and their final goal is to create an exhaustive list of documents to be delivered to the customer. By contrast journalists, analysts, and translators were polyglots and know the languages they are searching very well; their final goal is to use the retrieved information for writing (articles, reports, translations). However their knowledge of search strategies and text collections can be limited, sometimes really poor even thought they may have attended training courses.

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. As a consequence, the interface design was not influenced by their specific needs. In the following only journalists and translators are considered[3].

5.3 User Requirements

A list of user requirements derived through comparisons and discussions was produced. With respect to our initial idea (see 3.1) there are both differences and similarities. Generally people do not search for languages they do not know well, but can do in specific circumstances, e.g. the journalist

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

would use machine translation on authoritative online newspaper in other languages to double check an uncertain fact. No users performed periodic searches on the same topic, though translators may search for the same term several times in a month. Finally using many languages simultaneously to search was discovered to be an important feature not anticipated by the researchers. The main findings are listed below:

1. Users who know the language they are searching in (the majority) do so for other ultimate purposes (e.g. writing) 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 always use the most appropriate language they know for the task in hand, that is not always their native one.
4. English is used as a pivot to search other languages because of its dominance 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 and 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.

These findings have implications for the design of different aspects of the system: a) the user interface (1, 3, 5, 7), b) information retrieval functionality (2, 6, 6), and c) cross-language task (1, 3, 4, 6). The table below shows the possible impact of the findings the different components and suggests appropriate rectifying actions.

	Interface	**Information Retrieval**	**Query translation**
No query translation control	Hide query translation		High effectiveness
No result overview	Show ranked list first		
Multi-language search	Select many target languages	Support multi-collection search	Support multi-language translation
Multi-collection search		Support multi-collection search	
Dynamically swap languages	Source/target languages always ready	Dynamic swap among collections	Dynamic swap among languages
English as pivot	Allow un-translated search		Bypass the query translation
Multi language queries	Mark language term		Separate languages, translate when appropriate
Result filtering and sorting		Result categorization	
Result search	An appropriate panel	Boolean search on retrieved set	
Proper names			Algorithms for proper names translation
Phrases, compound names	Allow phrase search	Support phrase search	Algorithms for phrase translation
Term variation		Stemming, query expansion	Stemming, query expansion
User-centered dictionary	An appropriate panel	Allow search from dictionary selection	Support semi-automatic dictionary creation

Table 1. A summary of findings and their effect on Clarity components.

Although some findings may open new directions in CLIR research, for example how to properly translate phrases or compound names, current cross language technology (Pirkola et al. 2001) can largely satisfy these requirements. What is needed is a well thought-out use of the technology available based on a sound knowledge of the real context of use. What might seem complicated, i.e. support a multi-language query, is not a problem under certain conditions, particularly if the languages are made explicit by the user. Although the findings relate to the whole Clarity system, only the aspects relevant to the user interface are discussed below.

6. Designing A New User Interface

Following the analysis of user information seeking, we recognised some of our initial assumptions were wrong. This section reports on 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 simultaneously would have required a great deal of translation information to be displayed, the decision was taken to remove the query formulation and translation pane.

As users easily switch from one language to another, having the languages fixed in a separate window (the set-up pane) was not a good solution. In the new version (upper part of fig. 6) the search setting summary was removed whereas essential settings were always made available. Users can now move from one language to another, expand searches to further languages, or reduce the number of languages. 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; and translators swapped between source and target languages.

6.2 Result Presentation

Users expect to see a list of documents as result of a search and any other type of intermediate representation was considered annoying. This preference was recorded both in the participatory design sessions with the mock-ups and in the informal evaluations with Arctos that shows thumbnail documents with search terms highlighted (Ogden & Davis 2000). Even if users could quickly work out the meaning of these overview mechanisms, a glance at the top documents in the list would be enough to pick up the interesting bit. The result overview can be useful at a later stage, when and if a deeper investigation of the result is needed. Therefore this has been included as a supplementary panel in the final layout.

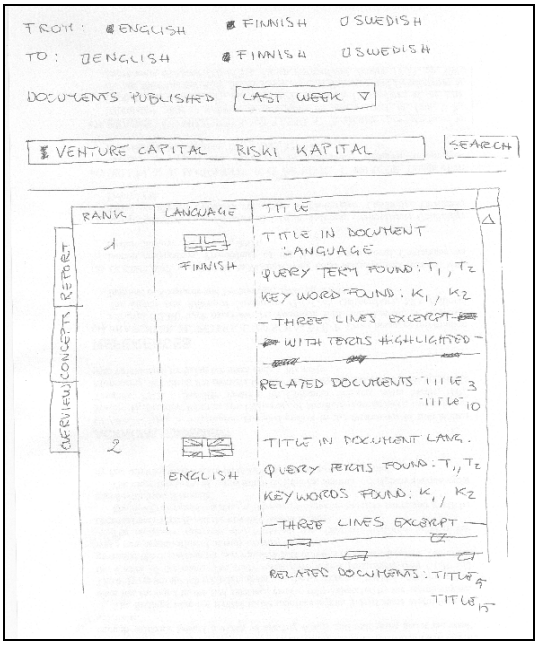


Figure 6. Mock-up of the revised interface

The ranked list was also extensively revised. In the new version (fig. 6, lower part), the basic display was simplified. Because the target 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 the MULINEX system (Capstick et al. 2000) can be reconsidered. Features related to the content such as concepts or style are kept separate and accessed by 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.

7. Interface Refinements

Figure 7a shows the first implementation of 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: tabs should be horizontal and at the top of a document.

- Alignment: document titles are more important than their attributes (i.e. rank position, language). Therefore it was decided to align the title on the left and rearrange the attributes.
- 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 7b, was the one planned for the user evaluation. Note that the best passage ("excerpt" in fig.7a) was not provided by the Clarity search core system at that time and was removed from the interface before the user test. Two graphic buttons were added: one to add a document to the working area ("add to keep list" in fig.7b) and one to see the query translation. Finally the user could select the number of documents to be retrieved in blocks of 20, 50, 100.

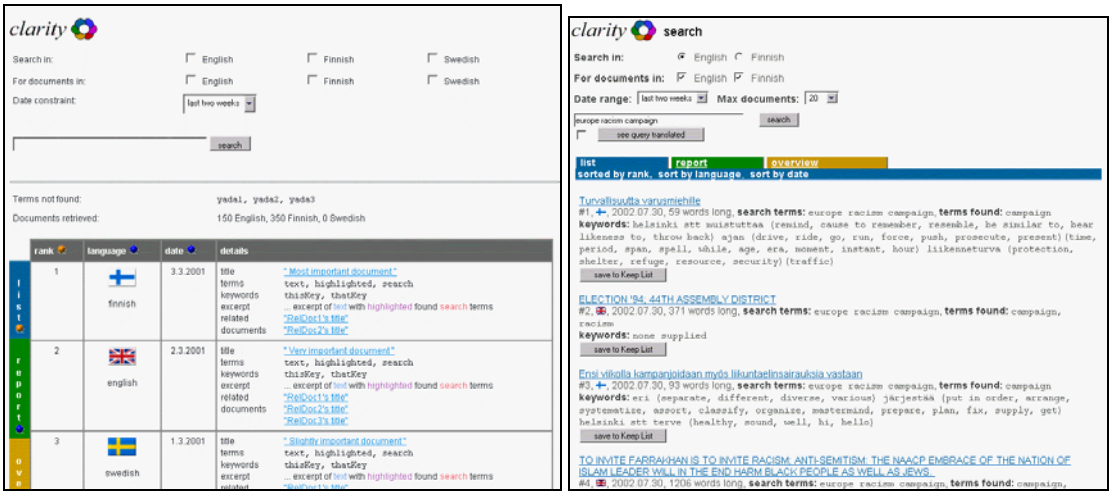


Fig. 7. a) The first implementation and b) the refined (tested) layout.

8. Using Empirical Evidence to Drive System Design

Compared to the usual perspective adopted for CLIR systems, the user requirements extensively changed the initial design and led to the decision of hiding the query translation mechanism in favour of a simpler layout and interaction. However this choice raises a set of fundamental questions on the proficiency of the CLIR and on the cognitive impact this complex task has on users. To test the system effectiveness and to better understand the cross-language searching task a set of user tests were undertaken. At that stage it was important to get ideas for the design, therefore several different and limited usability tests were preferred to a single, comprehensive user experiment.

8.1. Conditions, Interfaces and Users

The crucial point of "hiding" versus "showing" the query translation was tested as the main condition by all users:

- 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 steps CLIR), showed in figure 8a [4];
- the user inputs the query, the system translates the query and searches without any user intervention (1 step CLIR), showed in figure 8b.

Ten people participated in the tests: six monolingual subjects (English) searched Finnish documents; two polyglots (Finnish, first language, English second) and two monolingual subjects (English) performed multilingual searching with English and Finnish simultaneously.

[4] This layout is an approximation of the initial design, 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.

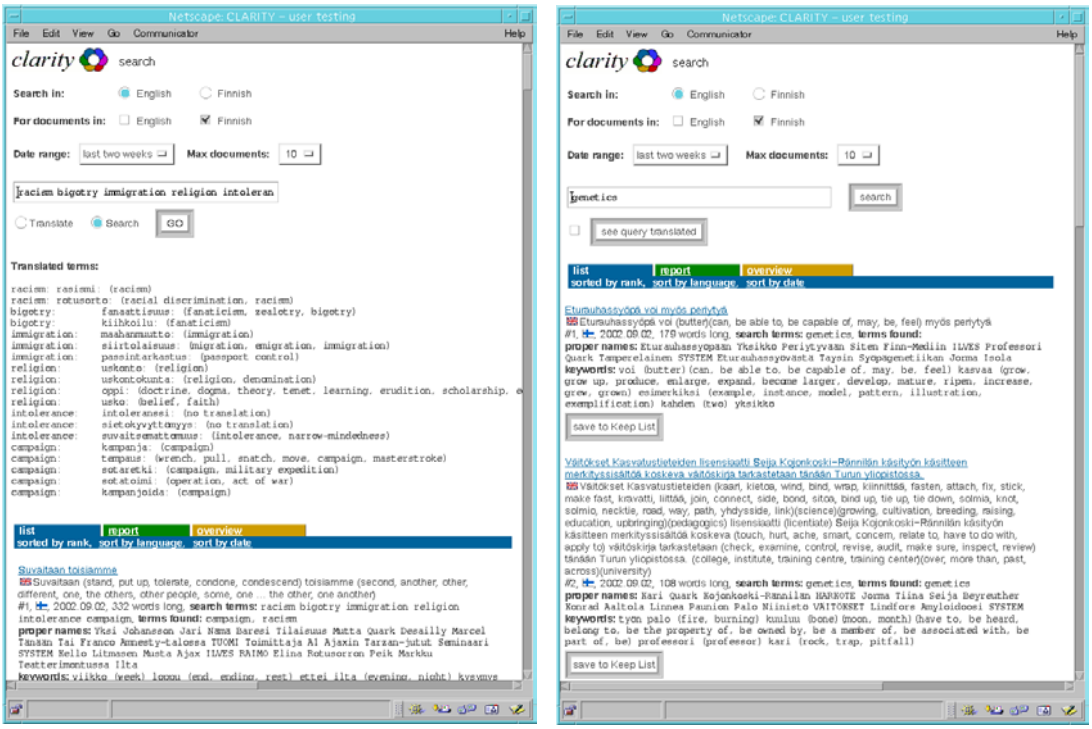


Figure 8. The Clarity layouts as tested in single language condition: a) 2 steps CLIR; b) 1 step CLIR.

The full Clarity prototype system used was physically distributed with a co-ordinating server and interface at the University of Sheffield, UK, and cross-language services (query translation and retrieval) at the University of Tampere, Finland, with communication facilitated by SOAP[5].

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. Participants were then provided with a scenario that described the task in hand (Borlund & Ingwersen 2000). The same training search was performed on each system, and then participants received four scenarios, two for each system. During the tests participants used both input interfaces and selected 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 overall test session. An automatic log recorded the queries submitted, the retrieved set, and the relevance judgements 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 and Oard's iCLEF guidelines (Gonzalo & Oard 2002).

8.2. Results and Discussion

The huge variation amongst users, tasks, and execution time made it impossible to gain a clear idea on which input interaction was more effective from a retrieval point of view (Petrelli et al. 2003). However by observing the users and analysing the log we gained a better understanding of the effect of displaying the translated query. The most important issued are discussed below.

8.2.1. Implication of the Query Translation Step

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. Almost all the observed subjects were puzzled when the system showed the query translation. They expected to see the result of the search and waited (for a while at times) for it to appear. Seeing the translation encouraged participants to reformulate the query before the search was run. By favouring generic query terms, such a strategy might limit the retrieval of relevant documents.

A similar phenomenon occurred when the user explicitly asked to see the query translation (by selecting the "see query translated" button, as shown in fig. 8b). In this condition, the translation was displayed but the search was automatically performed. Seeing multi-sense words led participants to

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

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.

We consider this result to be an important finding for cross-language research. Seeing the translation changes user searching behaviour and also potentially affects CLIR system performance. 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. The 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 to change the query or indeed why it should be changed. Showing the query translation may force the user to try more strategies than they would do so if they were left alone. The design could then adopt the more cognitively demanding solution (having the user check the query translation) if this would prove to be the most effective overall. A recent experiment by He et al. (2002) confirms this contradiction where users are 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 other hand changing the query before effectively running it may prevent the CLIR system from retrieving relevant documents. 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. Another user did not use "Alzheimer" because at query translation time the system showed that no translation was available. However if the term had been used, it would have effectively retrieved documents since un-translated words are passed to the retrieval mechanism as they are. Overall the data collected was not enough to clarify this issue and a new broader user test is planned to further explore this crucial CLIR question.

8.2.2. Names, Translations and Culture

Query formulation is always affected by user's knowledge. 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 English speaker searcher knew that Bobby Sand was a hunger striker, the system searched for Finnish translations of "policeman" and "beach".

Several participants used such searching by example strategy, mainly with proper names, often placed in inverted commas ("…"). Sometimes the translation was successful (e.g. Europe and Jew), others were not (e.g. Alzheimer). This type of failures is quite common in CLIR because it is a cross-cultural task and dictionaries or translations are limited in providing an appropriate context.

8.2.3 Language Knowledge Matters

Knowing the target language makes the search result more effective. Monolingual users were often unable to detect the relevant documents in the retrieved set even if many were available and very often judged as relevant documents that were not (Petrelli et al. 2003). This is due to the poor word-by-word translation adopted. However this problem cannot be avoided in the case of rare languages since more sophisticated tools and machine translation software are not available. Given the very high risk of misjudgement CLIR systems with weak translation support should not be suggested to monolingual users. Conversely the same users will be perfectly fine if presented with machine-translated documents (Bathie & Sanderson 2001).

Language knowledge also affected the user's general behaviour. Finnish speakers only read the title of the Finnish documents and did not open them to look at them more 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 keywords provided, while often they opened the English ones. This can be explained by the fact that Finnish for those users was a real barrier, thus looking at the document was not worth doing. Clearly the influence of the language in CLIR interaction warrants further investigation.

8.3 System and Interface Redesigned

From the Clarity design point of view the test results were very valuable. The insights derived from the user studies were confirmed, thus further supporting the design choices. Moreover the tests clearly showed which were the weakest points in need of attention.

The most frustrating aspect was the system response rate: It could take minutes to retrieve 20 documents, let alone 100. In spite of coding efforts to speed up the system, the target of retrieving in less than 10 seconds was not achievable. Thus it was proposed to retrieve documents 10 at a time. This design choice affected the interface as well as the retrieval mechanisms and reduced the response rate to around 5 seconds.

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. It was therefore decided to limit the translation of generic and polysemous words to only the most common terms. This will on the one hand present the query translation in a more compact way at the interface level and on the other hand make the searching more effective, since the distortion introduced by uncommon senses will be automatically removed.

Users expect the system to support a more sophisticated input than a simple list of words, which was all the system could accept. Participants tried to specify phrases using inverted commas, mandatory words using '+', and variations, e.g. "DNA", "D N A", "D. N. A.". Searching phrases, proper names, and un-translated words was considered mandatory for an effective CLIR system. Thus a grammar was introduced to communicate to Clarity which words had to be translated and which did not, as well as indicate which words should be searched as a phrase. This design choice would impact on both the searching system as well as the interaction by bypassing the query translation and having to introduce an additional search-window.

A few more findings affected the interface exclusively, as discussed in the following. Firstly, participants tried to disambiguate the query by using the Finnish word corresponding to the sense of interest. This made the system perform worse and frustrated the users. The new layout should allow users to simply deselect unwanted translations. Secondly, 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 an interface redesign is needed to make the sorting options clearer, the default for multilingual output should be by language. Thirdly, the working area (keep list) was well received, but the fact it was displayed in a separate window made the checking difficult for some users. In the new design a single window will encompass both the result and keep lists. Lastly, participants preferred not to see documents they had already viewed. This feature was particularly annoying due to the test system's slow response. Highlighting previously viewed documents could address this.

9. Conclusions and Future Work

The project reported in this paper demonstrated the value of user involvement throughout the design process. In the first place it dispelled false assumptions and misconceptions regarding user needs for CLIR and their motivation for searching resources in different languages. It became clear that searching behaviour depended not only on user goals or purpose for searching, but was also closely related to the language knowledge of individuals and the cognitive demands of the cross language task itself. Having gained a better understanding of the complexity of the cross-language task, the challenge for the system designer is how to best accommodate the different requirements with regard to both the system functionality and interface features.

The approach adopted here was to test two basic design conditions: a version that displayed the query translation and another that didn't. Whilst the jury is still out on whether or not it is best to hide or show the translated terms, the user evaluations provided some insight into the design trade-offs in building a system that meets user needs and is also effective. The weaknesses identified will now feed into another redesign cycle. However future evaluations will need to focus more closely on the tensions between the cognitive mechanisms and system efficiency for CLIR. In particular, there is the need to

address the three evaluation criteria of usability (namely efficiency, effectiveness and user satisfaction) in CLIR and explore the effect on different single or several target languages.

ACKNOWLEDGEMENT

Clarity is an EU 5th framework IST project (IST-2000-25310). 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 thank the partners for their collaboration and the people who volunteered as subjects. We are indebted to Heikki Keskustalo and Bemmu Sepponen from the University of Tampere for the promptness, patience, and help in setting-up the CLIR core module for the user tests.

REFERENCES

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.

Bathie Z. and Sanderson M. (2001) iCLEF at Sheffield. In C. Peters (eds.), Working notes for the CLEF 2001 Workshop (pp.215-217), Darmstat(Germany):ERCIM.

Beaulieu, M. (1997) Experiments on Interfaces to Support Query Expansion. Journal of Documentation, 53(1), 8-19.

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 (ACM SIGIR'96). (pp. 128-136), Zurich (Switzerland):ACM

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.

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

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

Cousin, S. B., Paepcke, A., Winograd, T., Bier, E. A. and Pier, K. (1997) The Digital Library Integrated Task Environment (DLITE). 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), :ACM.

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., Oard, D. (2002) The CLEF 2002 Interactive Track. In working notes for the CLEF 2001 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)

Hendry, D.G. & Harper, D.J. (1997) An Informal Information-Seeking Environment. In Journal of the American Society for Information Science, 48(11): 1037-1048

Hull, D.A., Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of the 19th annual international conference on Research and development in information retrieval (ACM SIGIR96) (49-57). Zurich(Switzerland):ACM.

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.

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

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., & Draper, S. eds. (1986) User centered system design. New perspectives on human-computer interaction. Hillsdale, N. J.: Lawrence Erlbaum Ass.

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

Oard, D., & Gonzalo J. (2001) The CLEF 2001 Interactive Track. In working notes for the CLEF 2001 Workshop, 3 September, Darmstadt, Germany, 203-214.

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

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., Demetriou G., Herring P., Beaulieu M., & Sanderson M. (2003). Exploring the Effect of Query Translation when Searching Cross-Language. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck eds. Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign, Springer Lecture Notes in Computer Science LNCS 2785.

Pirkola, A., Hedlund, T., Keskustalo, A. & Jarvelin, K. (2001). Dictionary-based cross-language information retrieval: problems, methods, and research findings. 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, (pp. 475-489).

Rosson, M. B., & Carroll, J. M. (2002). Usability Engineering – Scenario-Based Development of Human-Computer Interaction. Morgan Kaufmann.

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

Sanderson, M. 7 Croft, W.B. (1999). Deriving concept hierarchies from text. Proceedings of the 22nd annual international conference on Research and development in information retrieval (ACM SIGIR99), (pp. 206-213), Berkley(USA): ACM.

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

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