On the role of user-centred evaluation in the advancement of interactive information retrieval

PETRELLI, Daniela <http://orcid.org/0000-0003-4103-3565>

This is a preprint of a paper accepted for publication in Information Processing & Management - special issue on "User-Centred Evaluation of IR Systems", Pia Borlund & Ian Ruthven eds.

On the Role of User-Centred Evaluation in the Advancement of Interactive Information Retrieval

Daniela Petrelli
Department of Information Studies
University of Sheffield
Regent Court
211 Portobello Street
Sheffield S1 4DP, UK
d.petrelli@shef.ac.uk

Abstract

This paper discusses the role of user-centred evaluations as an essential method for researching interactive information retrieval. It draws mainly on the work carried out during the Clarity Project where different user-centred evaluations were used during the lifecycle of a cross-language information retrieval system. The iterative testing was not only instrumental to the development of a usable system, but it enhanced our knowledge of the potential, impact, and actual use of cross-language information retrieval technology. Indeed the role of the user evaluation was dual: by testing a specific prototype it was possible to gain a micro-view and assess the effectiveness of each component of the complex system; by cumulating the result of all the evaluations (in total 43 people were involved) it was possible to build a macro-view of how cross-language retrieval would impact on users and their tasks. By showing the richness of results that can be acquired, this paper aims at stimulating researchers into considering user-centred evaluations as a flexible, adaptable and comprehensive technique for investigating non-traditional information access systems.

Keywords

Interaction design, iterative user evaluation, empirical studies, cross-language information retrieval.

1. Introduction

Very often user evaluations are done in IR at the end of the system development process. However, this approach is likely to have a limited impact on the system design since the main choices have been made and a lot of implementation effort has already been spent. This type of evaluation undertaken at the end of a design project is referred to as a *summative evaluation*, since it sums up the work done so far. However, user evaluations can be run at any time during the design process. These are *formative evaluations* and help in finding out critical points in the interaction and thus contribute to the transformation of the design in progress. The goal of these iterative evaluations is to verify if the interaction is as expected, find out where the problems are, understand what is wrong and how it may be addressed (Dumas & Radish 1998): they are an essential element in system design as they keep the focus on the user (Preece et al. 2002).

In the area of Interactive Information Retrieval (IIR), the iteration of design and evaluation has been identified as key to achieve effective systems and the way to avoid the "user-centered design paradox" (Marchionini 1995): "We cannot discover how users can best work with systems until the systems are built, yet we should build systems based on knowledge of users and how they work." "The solution [to the user-centered paradox] has been to design iteratively, conducting usability studies of prototypes and revising the system, over time." (Marchionini 1995, p. 75).

Despite the recognized importance, there is a general lack of studies that show how design of IIR systems evolved and explain the rationale for the choices made. This could be because formative evaluations are not often considered a research instrument rather just a tool for practitioners, a way to produce an IR system that works in the real world (e.g., Schusteritsch et al 2005). For research purposes other methods are preferred, e.g. qualitative experiments for information seeking and behaviour studies (e.g., Kelly & Belkin 2004, Talja 2002), quantitative studies of log records (e.g., Anick 2003, Ozmutlu et al. 2004) or comparative tests for user interface features (e.g., Hughes et al 2003, Koenemann & Belkin 1996). These

methods allow one to study in depth a fraction of the user-system interaction, but fail in capturing much of the context (the system, the user, or the environment respectively) and do not provide a holistic understanding of the phenomenon under study. When studying new areas of interactive information retrieval (e.g. cross language, multimedia, or personal information retrieval), a tool that supports an explorative and inquisitive approach is more promising than one aiming at a definitive answer.

This paper discusses the role of user-centred evaluations as an essential method for researching interactive information retrieval (IIR). It draws mainly on the work carried out during the Clarity Project where different user-centred evaluations were used during the lifecycle of a cross-language information retrieval (CLIR) system. The iterative testing was not only instrumental to the development of a usable system, but more importantly, it enhanced our knowledge of the potential, impact and actual use of CLIR technology. By showing the richness of results that can be acquired, this paper aims at stimulating researchers into considering user-centred evaluations as a flexible, adaptable and comprehensive technique for investigating non-traditional information access systems. A set of related evaluations allows understanding of the evolution of a certain interaction design, i.e. to understand the *why* and not only the *what*. In the case reported here for example, the design choice for the interaction was explicitly done against user preferences. By understanding why users preferred one solution against another, it was possible to find a compromise that satisfied users and exploited the system capabilities.

2. Interactive IR Evaluation

2.1. Ecological vs. Controlled

The system-oriented laboratory-based IR evaluation framework has been challenged in the past, particularly with respect to the lack of user involvement and attention to interaction. Critiques include: lack of insight into the user-system interaction (Robertson & Hancock-Beaulieu 1992), narrow focus on the system at the expense of the searcher and the context (Saracevic 1995) and a disregard for iterative and exploratory retrieval (Draper & Dunlop 1997). Ingwersen and Järvelin (2005, p. 7) go further and list ten objections to the laboratory-based evaluation framework including limitations of precision and recall in representing a successful interaction, and a leaning toward average results to the detriment of a deeper understanding.

The need for a distinct and broader evaluation framework for Interactive IR (IIR) was recognized through the many years of the interactive TREC track and a lot of effort was spent to move user-evaluation from a system-orientated perspective to a more realistic user-orientated one. As observed by Over (2001), the need for a broad consensus among interactive TREC participants led to a generic framework focused on strictly controlled laboratory user evaluations. This centralized control of the experimental design was essential for result generalization but hampered a more operational and ecological approach. Studying the interaction in natural settings (e.g. offices, libraries) with real users each with a well-defined information need, is the ideal way to investigate reality. Indeed, the lack of realism in controlled user evaluation has been criticized and the need to perform more realistic tasks has been strongly advocated (Robertson & Hancock-Beaulieu 1992, Su 1992). Unfortunately an ecological approach fails to collect homogeneous data indispensable to address usability issues. To understand how serious a problem it really is, experimenters have to know how many people encountered specific difficulties. However, if users search different topics (as in the ecological setting) the outcomes are not comparable. When the focus is a quantitative study, conducting an experiment under controlled conditions is essential and a laboratory is the best place.

Controlled tests and ecological observations are the two extremes of the user evaluation spectrum, in between the two are many degrees of intermediate solutions. Evaluations run periodically during a project lifecycle allow for a progressive move from strictly controlled laboratory test to more relaxed conditions resembling naturalistic observations. In the early stages of system design it is far more important to understand why the system is not working properly, while when a stable prototype is available the goal of a broader investigation is to assess the impact of the designed system on the user in their natural context. Some forms of user-centred evaluation can also be used in the very early stage of the project when ideas are generated to more effectively identify user requirements. From this point of view periodical user evaluations are a kind of longitudinal study done on the same evolving system: they allow investigating if and how patterns of user behaviours change or stay when the interaction changes inside a given framework. As a research method longitudinal study allows for the exploration of alternatives but similar solutions, and supports a steadier accumulation of knowledge on how users would interact with a certain technology besides the specific implementation. This incremental and explorative approach is not widely adopted in the IIR community: the work of Nick Belkin and colleagues at Rutgers University is an exception and not the rule. In three years of Interactive TREC (2001-2003) they investigated several interface features to discover the effects on query length (Belkin et al. 2003, Belkin et al. 2002).

2.2. Operational vs. Hypothesis-based

User evaluation of innovative IIR systems is often seen as a part of the software development process. The user test is run to measure how the system performs when put in use (as in Dumais et al. 2003). This operational evaluation is generally an approximation of real-life conditions: a system retrieves documents in real time; a fully developed user interface is provided to the user to interact with; a task is given to the user to be carried out. The performance analysis can provide good insight into the effectiveness of the new technology as well as its impact on users and their satisfaction. However, its potential to advance the knowledge of interactive IR is rarely exploited and the result offered to the research community is often just the assessment of the developed system.

Another form of IIR user evaluation is the hypothesis-based: a proposition is formulated in advance and the evaluation is set up to confirm or refute the underlined theory. In the most of the cases two interface conditions are compared to test which one works best. There is no constraint to reality: the system can effectively retrieve (Sav et al. 2006) or be just a simulation (Dumais et al. 2001); the user interface can support the accomplishment of a task (McDonald and Tait 2003) or just collect user's answers (e.g. relevance judgement) (Oard et al 2004). The research is quantitative and targeted to evaluate a single hypothesis. Result generalization is possible, but depends on the question and the statistical significance of the results. The utility of this type of evaluation is in the evidence they offer in terms of which feature works best: designers of interactive IR should use the outcome of these types of experiments when deciding on the composition of the interface and the dynamic of the interaction.

Though different in form, these two types of evaluations should not be considered antithetic; neither should the second be considered more research than the first one. Indeed from the point of view of the advancement of interactive IR research both have advantages: the first is more innovative (i.e. a new IIR system) and offers a broader view, the second is highly focussed and more reliable. A synergistic use of the two types of evaluation would allow a better investigation of innovative interactive IR: operational evaluation provides a generic understanding, hypothesis-based evaluation provides empirical evidence.

Evaluations are the core of the user-centred design approach where iteration of design and evaluation support the understanding of the context and quality of use as well as the testing of the system performance. The next section outline the user-centred approach for information access and discusses how the different evaluations (ecological or controlled, operational or hypothesis based) fit in the general framework. Section 4 reports on the series of user-centred evaluations (some operational oriented, others hypothesis-oriented) periodically carried out during the design and development of a cross-language retrieval system. The aim is to discuss how the different user-centred evaluations were used to direct the design and the rationale for the choices made. All evaluations served to increase the level of understanding of the human-CLIR interaction.

3. Roles of User-Centred Evaluation in IIR

Until recently, the three areas of IIR, systemic IR and information seeking have for the most part followed their own research agenda. Ingwersen and Järvelin (2005, p.315) have classified the objectives of information seeking research, interactive IR research and system IR research and observed how little overlap among the three disciplines has been accomplished so far. In their view, information seeking research has been successful so far in developing a theoretical understanding of the seeking process and in providing an empirical understanding of the underlying phenomena, but has fallen short in supporting information management and system design. The reason for this lies in the focus of research, which has generally excluded the system (Ingwersen and Järvelin 2005).

In contrast, when the IR system is at the centre of the investigation and the interaction is the focus of the research, the users are involved, but the sample is generally opportunistic (e.g. students recruited in the institution) and the context is not real. This de-contextualization allows basic interaction factors to be measured (e.g. completion time, number of relevant document retrieved) but fails in capturing the quality of information with respect to its use or the process that triggered the search activity. The dichotomy is reflected in the opinion that each other's results are of limited use: IR research sees information seeking results as short of practical utility ("unusable academic exercise" in Ingwersen and Järvelin terms, p.2); information seeking sees IR research as lacking in understanding and abstraction ("too narrowly bound with technology" in Ingwersen and Järvelin terms, p.2).

To overcome this separation, Ingwersen and Järvelin advocate a holistic perspective and propose a cognitive framework of nested contexts of information retrieval, information seeking and work/interest (Ingwersen and Järvelin 2005, p. 322). They list a set of dimensions (e.g. natural work/search task, actor characteristics, etc.) and a number of variables for each of them, and discuss how both IR research and information seeking studies should enlarge their perspective to include those aspects of the context.

However complexity increases with the number of aspects and only simple cases can be fully investigated in a single evaluation. An iterative framework can support the creation of a holistic view by composing and cumulating the knowledge derived by each single evaluation. As in the case of Clarity discussed in section 4, user-centred evaluation can function as a deductive or inductive tool depending on the set up and the time in the system design life it is performed, and supports the researcher in moving between the theory (deductive approach, to prove a theory or confirm a hypothesis) and the empirical data (inductive approach, observations and intuitions leading to concept formation) in an abductive way (Manson 2002).

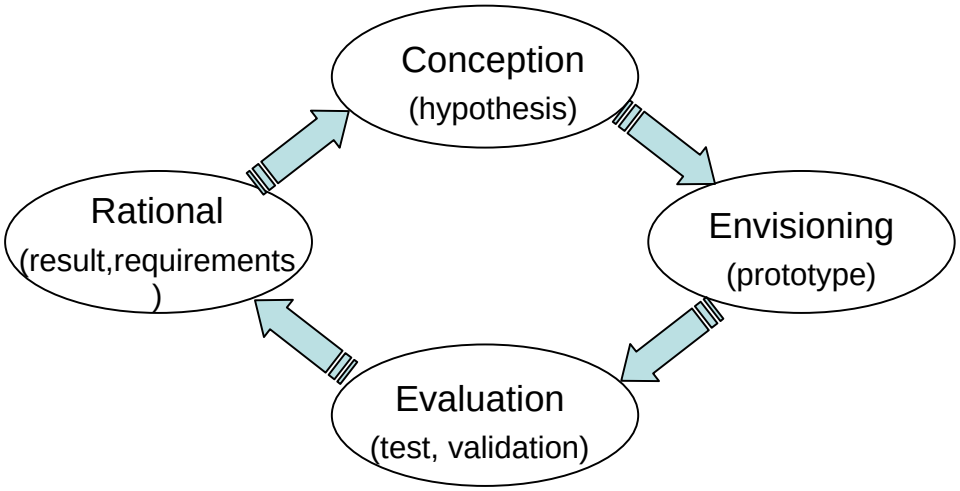

Figure 1. The design-evaluation cycle.

Figure 1 shows the design-evaluation cycle that is the foundation of a user-centred approach. There is no prescribed starting point; what is core is that each phase impacts on the following. An iterative user-centred design as the one implemented in Clarity is articulated in four phases, each encompassing at least one cycle:

1. *First Cycle: Understanding Users, Tasks and Environment*

   The first cycle should provide an understanding of the broad context of use (i.e. study the "work task", Ingwersen and Järvelin 2005, p. 322): information behaviour research is likely to match this stage and should be used to inspire the design. Studies centred on the individual, their motivations and their environment fit this initial cycle and are instrumental to gain a generic understanding, can support the selection of realistic tasks and feed the system design phase (Whittaker et al. 2000). The informal test of an existing CLIR system and the observational study done in Clarity (section X.1) are example of the inductive research appropriate to this stage. Indeed any implementation prototype created as a proof of concept (Hounde and Hill 1997) can be used to start an iterative investigation and point out potential problems that could became the core of deeper investigations.

2. *Second Cycle: Testing Ideas with Low-Fidelity Prototyping*

   Several studies have shown that the data collected when low-fidelity prototypes (i.e. paper mock-ups) are used is as reliable as that collected with an actual prototype (Virzi et al 1996, Catani & Biers 1998, Sefelin et al. 2003). Although rarely used in the design of IIR systems, paper mock-ups can be used at a very early stage to validate design ideas. In Clarity paper mock-ups were validated in participatory design sessions (section X.1).

3. *Third Cycle: Advancing System Design with High-Fidelity Prototyping*

   As the design of the system progresses, evaluations are likely to include only part of the context of use and focus on the IIR and IR aspects, namely interaction and retrieval. Controlled evaluations that make use of real tasks and representative users are one example. Evaluations of this kind help the design progress by cumulating knowledge and understanding on how the system performs under realistic conditions and how users perceive it. More variables (in the

sense of Ingwersen and Järvelin above, i.e. different tasks, users, etc.) can be introduced and tested when previous questions have been answered and the prototype consolidated. Sections 4.2 and 4.3 describe two controlled evaluations (one hypothesis-based one more operational) that motivated the evolution of Clarity design. User evaluations done at this stage are the most important for the progress of the system design and should inspect the different software components as well as the different dimensions of the quality of use (this issue is further discussed in section 5).

4. *Fourth Cycle: Overall Validation*

At this point in the project lifecycle the needed functionalities are in place (e.g. fast completion, appropriate number of translations) and the evaluation should move from a controlled setting into the real context of use and should address the quality of use of the designed system into a real work environment. Section 4.5 reports Clarity final evaluation that included ecological aspects (work context, final users and individual tasks). At the end of the iteration the user satisfaction can become the most important parameter (Su 1992), possibly not the only one its value is strongly influenced by personality and attitude (i.e. computer anxiety) (Johnson 2005).

The next section provides more details on how the five user evaluations performed in the Clarity project advanced out knowledge of user-CLIR system interaction as well as the system design.

4. User Evaluation in Action: A Case for Support

Clarity[1] was a EU-funded project aimed at creating an interactive cross-language information retrieval system (CLIR) for rare languages, namely Finnish, Swedish, Latvian and Lithuanian. The purpose of CLIR is to allow a user to search text documents written in a language (destination language) using a different language for the query (source language). For the retrieveal to succeed, the query or the documents must be translated into the other language (or both into a third one). Translating the whole document collection from the destination language into the source language requires specific machine translation software; the translation of the query instead can be done using machine translation or a simpler bilingual dictionary (Oard 1998). Which method can be applied might not be related to the effectiveness of the technique alone. The translation of the whole collection requires knowing in advance which languages the user could use to query; this could present problems of storage and updating.

Although machine translation is becoming more and more popular, it is still limited to the most widespread languages, for many others the only electronic resource available is a dictionary and the only option is the translation of the query. This was the case of the languages Clarity was dealing with. Therefore, in the following analysis all the evaluations refer to the mechanism of translating the query from the source language to the destination language and use the translated query to search.

From the user point of view, the process of cross language interactive information retrieval is articulated into several steps: the user inputs a query in a source language; the system translates the query; the translated query could then be displayed to the user for validation before being used to search a collection of documents written in the target language (different from the source one); finally, the result could be translated into the user language before being displayed. Different user evaluations were run during the lifecycle of the project:

1) to form an initial understanding of the user-CLIR interaction in order to foresee potential problems with the envisaged interface (informal evaluation of paper mock-ups and existing CLIR system during the requirement analysis);
2) to base the interaction design on empirical evidence (two formative evaluations);
3) to test the final design and investigate the potential of the developed technology with actual users (summative evaluation).

A detailed report about all three evaluations is beyond the scope of this paper and hence only contextually significant examples are presented. A more detailed analysis is in previously published papers (Petrelli et al 2002, Petrelli et al 2004, Petrelli et al 2006).

4.1 Testing Ideas: Paper Mock-ups and Informal Evaluations

The early stages in the design of a new system are devoted to the definition of requirements, what the system will do and how. In a system-centred view, the requirements are the functionalities the system will provide. Formalism exists (e.g. UML) to precisely define use-cases that describe how the system will react to user's actions and, although the focus is on software engineering, attention is paid to the user and their social context (Goguen & Jirotka 1994, van Lamsweerde 2000).

[1] http://www.dcs.shef.ac.uk/nlp/clarity/

As discussed in Section 3, a user-centred approach starts with a cycle of observing and understanding the user context in order to define an appropriate set of user requirements for an effective design (Ackos & Redish 1998). Starting with the user point of view, the requirement analysis can suggest a different system than the one initially envisaged by the researchers and can therefore help in identifying and removing misconception before any final decision is taken.

Both system and user requirements are needed: system requirements provide what is technologically feasible, user requirements what is actually useful. The role of the interaction designer is to mediate between the two (often contrasting) points of view and come out with a compromise that exploits the technology in a way that users would find valuable. This mediation takes place during the conceptual design: an abstract picture of what information and functions are needed for the system to achieve its purpose is outlined together with a conceptualization of the envisaged solution and how that conceptualization will be communicated to people (Benyon et al. 2005). The conceptual design feeds the physical design: interface and interaction details are fully specified and linked to internal functionalities (Benyon et al. 2005).

4.1.1 Goal and Set-up

Clarity user requirements were the outcome of a field study (Petrelli & Beaulieu 2002, Petrelli et al. 2002). Five complementary techniques were used to collect data and compose a picture as extended and realistic as possible[2]: contextual inquiries, interviews and questionnaires allowed users and tasks to be classified, while informal evaluations of an existing CLIR system and participatory design sessions were used to test Clarity preliminary ideas against potential users. Informal evaluations and participatory designs can be seen as kinds of user-centred evaluations and are therefore discussed further.

In line with the few interactive CLIR systems available at that time (Ogden et al. 1999, Ogden & Davis 2000, Capstick et al. 2000), the Clarity interaction was split into two phases: query translation checked by the user followed by actual search. Key points of the interaction were captured in paper mock-ups. In the participatory design sessions, scenarios of use describing a user and his/her information needs were discussed with participants and used to explore the interaction with the system visualized by the mock-ups. The use of paper mock-ups instead of a working prototype to test initial ideas had the advantage of focussing the discussion on the functionalities instead of the layout; working with paper mock-ups users felt free to actively contribute by proposing (i.e. drawing) a different interface that better reflected their view and was not perceived as being dismissive of the work already done.

In the informal user evaluation, participants were observed trying out ARCTOS[3] (Ogden & Davis 2000), a CLIR system which was available on-line at the time of study. Although it could be set up as a formal evaluation (i.e. with tasks to accomplish and measures taken), this test was left informal and user directed. The intent was to observe how users would naturally interact with a CLIR system that shared some design rational with ours. Participants were questioned during the interaction in order to gain a better understanding of what was in their mind.

4.1.2 Data Analysis and Results

Both studies were videotaped. The analysis was qualitative and based on observations of participant's behaviour; users' activities were decomposed using six dimensions: goals, tasks, acts, community, practices and procedures, opinions and suggestions. Affinity diagrams (Ackos & Redish 1999) were created to cluster users' activities around few common and recognizable phenomena. Users were then classified respect to their ability, attitudes and expectations; requirements for each user class and the whole community were listed.

The two evaluation sessions pointed out potential problems with the anticipated user-CLARITY interaction. Users were not interested in controlling (or did not know how to control) the query translation step, nor were they interested in graphical visualisations of the global result. Instead, a simple mechanism of typing in the query and receiving back the list of relevant documents was expected.

A tension emerged between CLIR common practice (i.e. display the query translation for user validation) and user expectations (i.e. display the search result). Instead of choosing one of the two directions, a comparative user test was set up to investigate the two conditions in a more formal setting.

4.1.3 Lesson Learnt: The Value of Informal User-Evaluation

Trying out our ideas with users at a very early stage in the project design was essential to raise our awareness of potential problems in adopting the user-supervised solution as the interaction mode. The use of paper mock-ups and scenarios

[2] These are just a few of the many techniques that can be used to collect user requirements (Ackos & Redish 1999).

[3] Screenshots of the ARCTOS system are available at http://crl.nmsu.edu/~ogden/i-clir/cltr-interactive/arctos/page1.html (accessed 23.3.2006).

supported the participants in getting a grip on what the system would do and the low-technology setting was ideal to allow them to actively contribute.

The informal evaluation with the ARCTOS system was equally informative: By observing participants struggling with the query translation and the result display we became aware of how unfamiliar users would be with basic concepts of cross-language retrieval.

4.2 Learning by Doing: User Evaluation as an Exploration Tool

4.2.1 Goal, Set-up, and Data

A comparative user evaluation was undertaken to empirically investigate the two approaches (Petrelli et al 2004):

- *Supervised*: derived from the CLIR/IR literature; this required the user to check the translation before the search was done;
- *Delegated*: derived from the requirements analysis; this required the user to only input the query, then the system would translate it and search without any user intervention.

The experimental design followed the CLEF[4] framework (Gonzalo & Oard, 2002): a collection of newspaper articles, 4 topics (+ 1 for training) with their corresponding relevance assessment were used. 6 English native speakers with no knowledge of Finnish searched Finnish documents; in addition to a training task, they performed 4 tasks, 2 on each system. Users were given a simulated task (Borlund 2000) derived from the CLEF 2002 topics and were asked to retrieve relevant documents; documents judged as relevant were saved in a ad-hoc list. The total time of the experiment was 3 hours. Queries and relevance judgements were logged and time-stamped; questionnaires on user profiles and satisfaction were collected; observations were made whenever possible.

The data was not consistent enough to statistically select one interaction as the best and another test (3.3) was run later for this purpose. However the data was analysed through a qualitative inspection of the user's actions. All the user's queries were compared to the query translation (if presented to the user), the result displayed and the follow-up query. This allowed patterns of behaviours to be identified, as discussed below. The relevant judgements were analysed to assess the effectiveness of the result display: the documents selected by the user as relevant were compared with the list provided by CLEF in order to identify correct selections; false positive (documents judged as relevant by the user but that were not relevant in CLEF assessment) and false negative (documents not selected as relevant but that were listed as relevant in the CLEF assessment) were considered as well. System weaknesses emerged in both user interface and system functionalities.

4.2.2. Results and Changes

The user's perception was most affected by the speed of the system: every single user complained about the fact that the system sometimes needed minutes to return results. The system's architecture was redesigned and strategies were adopted to make the system more efficient (e.g. pre-translation of the titles). A response time of 5 seconds[5] was set as usability target in further tests.

The second weakness was the number of translations used for polysemic words. All the senses were included which made the search inefficient and the user confused as when, for example, the Finnish for "golf pitch" was proposed as the translation of "green". The number of translations was then reduced to the three most common. This also greatly simplified the result display as titles and keywords were translated using the same mechanism.

Seeing the query translation affected the whole interaction. The analysis of logs showed a tendency to change the query before the search if the translation included ambiguous terms. A user started with "green power" but ended searching with the non-ambiguous query "wind turbine", thus potentially missing relevant documents.

Proper names were widely used by participants but badly managed by the system. Some names where in the dictionary (e.g. Europe) thus were translated, others where not (e.g. Alzheimer), and others were wrongly translated (e.g. "Bobby Sands" a famous hunger striker was translated into the Finnish equivalent of "policeman beach"). Fuzzy name translation (Pirkola et al. 2003) would probably resolve the non-translation but not the misleading one. A new feature was introduced to allow the user to mark terms which must not be translated.

---

[4] CLEF stands for Cross Language Evaluation Forum and is the annual evaluation campaign for research on cross language information retrieval for European-based languages (http://www.clef-campaign.org/ accessed 10.7.2006).

[5] Jacob Nielsen reports (pp.44) 10 seconds as the limit for keeping the user's attention focussed on the interaction and 1 second as the limit for the user's flow of thought to remain uninterrupted (Nielsen 1999). By setting the response time to 5 seconds we were confident the interaction would be fluid enough to be successful.

An overwhelming preference for the Delegated mode (70%) over the Supervised one (15%) emerged from the questionnaires thus reinforcing what found in the field study (15% were neutral).

4.2.3 Lesson Learnt: The Value of Qualitative Analysis

The high variability in the data prevented any statistical analysis. Variability was due to task feasibility, user's search skills and tiredness:

- Tasks: one task was unfeasible (i.e. no user retrieved any document) and for another there was a very low success rate. Although being in the worst condition stimulated users into trying new strategies (which was positive for qualitative studies), the amount of data collected was affected by a nearly 40% reduction. Moreover some users can experience high level of frustration[6].
- Users: search attitude was measured by a self-rating questionnaire that indicated participants formed a homogeneous group, but the user's actual engagement with the search task (e.g. number of queries issued, number of different terms used) and the user's search effectiveness was disparate. A homogeneous group of users is needed to collect consistent data, essential in the early phases when the system is in evolution as its performance is the object of the study: variation in the performance can then be attributed to the different system design and not to the different users' capabilities. When a consolidated prototype is available, different classes of users can be involved to test the system under different conditions and detect previously uncovered problems (Caulton 2001).
- Tiredness: the engagement with the search tasks dropped steeply in the last two tasks for two of the six users[7].

These issues were taken into account when the next user evaluation was designed and effort was spent to mitigate the negative effects. By controlling task feasibility, user's ability, and evaluation duration more closely, we hoped to collect more reliable data that could be statistically analysed.

The data collected was very rich in terms of user behaviours: by analysing user's interactions we were able to detect pattern of behaviour that negatively impacted upon the system effectiveness, i.e. the focussing of the query after having seen the translation and before the search. Through the qualitative analysis technical limitations were also detected, e.g. the weak translation of proper names. Faults and weaknesses gave suggestions of what needed to be changed to make the interaction more successful. A statistical analysis would not provide such a rich set of inspiring examples of actual CLIR use that helped in re-focussing the design.

4.3 Directing Design: Controlled User Evaluation

4.3.1 Goal, Set-up, and Data

The new Clarity prototype took advantage of the changes made after the previous evaluation and was tested in a second formative evaluation (Levin & Petrelli 2003). The same text collections and relevance assessments as in the previous evaluation were used but a different set-up was planned to reduce data variability. The experiment was shortened with only 1 task per system; the 2 chosen were those with the highest success rate in the previous test across all users. Participants were 8 Finnish and 8 Swedish native speakers (bilingual with English) who tested 4 language pairs (En→Fi; Fi→En; En→Sw; Sw→En); subjects were screened through a practical task to confirm their search skills (participants used Google to search for "jaguar" -the animal, not the car- we observed their skills in generating new terms and disambiguate the query). More sophisticated tests can be set up to investigate users' ability (Saracevic et al. 1988), however to recruit people with similar searching skills or, at least, to exclude the worst, this simple screening was considered enough.
The exhaustive data collection encompassed: i) full log of the time-stamped interaction (e.g. queries, selection on the interface, documents opened, etc.); ii) videos of user's actions and comments; iii) questionnaires and interviews.

4.3.2 Data Analysis and Results

The data collected was rich in both objective and subjective measures; *objective* measures pertain to facts (like the number of terms and queries issued by each participant or completion time) and were recorded in logs, while *subjective* measures pertain to users' opinions (if they liked or disliked something) and were collected in questionnaires and interviews. Data was analysed both quantitatively and qualitatively. *Quantitative analysis* used aggregated values of objective measures to

[6] Participants can show real discomfort if they feel they are performing badly independently from the reassurance of the experimenter. During the evaluation of an image retrieval system two participants out of eight asked how the others had performed and did not want to give up the task even if the time had elapsed.

[7] Other two were consistently effective and two were consistently unsuccessful during the whole test.

determine system efficiency and effectiveness while subjective measures were used for user satisfaction. Observation and interviews were used in *qualitative analysis* to inspect each interaction and scrutinize behaviours.

The interaction was measured with respect to *efficiency*, *effectiveness* and *user satisfaction* (measures further discussed in 5.2). Efficiency was measured in time needed to get an answer from the system and effort spent. The threshold of 5 seconds was generally kept and no user complained about system speed. The effort spent was calculated in terms of number of user's actions. The mean of queries issued in the Delegated Mode (DM) was higher than in the Supervised one, but the difference was not statistically significant. However, the Supervised Mode (SM) offered the user the possibility of deselecting translated terms. Then the number of queries was used as measure of engagement with the DM interface, while for the SM the measure includes both the number of queries and the number of deselected terms. A paired-samples t-test was conducted: There was a statistically significant increase in the engagement from Delegated (M=6.23, SD=3.44) to Supervised (M=9.62, SD=5.05), $t(12)=-4.58$, $p<.001$. Indeed the possibility of deselecting terms was central as all the users deselected at least one sense (and up to 6) from those offered by the Supervised interface. The number of de-selections depended upon the words used in the context of the search task. In summary, the users interacted more with the Supervised mode, but produced less queries.

To assess the overall effectiveness of each interaction mode in supporting query formulation, average precision and recall (P & R) measures were calculated. The measurement took place at display time, before the users bookmarked the relevant documents. In other words, as for the batch approach, the output of the search was used to calculate P & R. This was done to avoid biasing the objective measure of effectiveness with the variability inherent in a subjective relevance judgement (Mizzaro 1997). In this way precision and recall measure the effectiveness of the query formulation step in isolation from the rest of the interaction. User relevance judgement was used to assess the effectiveness of the result display.

Although SM performed better (P= 0.206, R= 0.473) than DM (P= 0.167, R= 0.418), the differences were minimal and not statistically significant when a paired-samples t-test was applied. However, such small difference is still meaningful from a user point of view as it corresponds to at least one more relevant document being retrieved out of the 12-17 available in the collection, that is to say a 6 to 8% increase.

Users' relevance judgement was used to assess the effectiveness of the output display alone. The assumption was that a good presentation would allow the user to select a high percentage of the relevant documents retrieved and displayed. The portion of relevant documents correctly identified out of the set of the relevant documents retrieved was used. This can be seen as a sort of precision calculated on the basis of the retrieved set and not in relation to the whole document collection. The rationale behind this is that if relevant documents were not retrieved, the users had no chance to select them, therefore this distinguishes the effectiveness of the retrieval mechanism and effectiveness of the display. As the display was equal in both systems the result was cumulated giving a precision at display time of 0.57; this was not a high performance considering users were native or fluent in the target language and can be partially explained by the behaviour of judging relevance by the title, as discussed below in 4.3.3.

Users marginally preferred DM (Delegated=50%, Supervised=37.5%, Neutral=12.5%) but the difference was small and decreased greatly from the previous test (Delegated=70%, Supervised=15%, Neutral=15%). Interviews allowed the interpretation of the questionnaire results. Users' dislike for SM was related to a perceived slowness or the unnecessary obligatory step of checking the query translation and not to interaction complexity. Many users favourably commented on their increased control of the system and the inspiration for new terms which occurred in checking the translation.

Once again a contrast emerged from what the user preferred (Delegated) and what was more effective (Supervised). However in light of the insight offered by the interviews, the final interface is a compromise between Supervised and Delegated: the final interface automatically translates and searches (Delegated), and then the query translation is displayed on top of the result list (Supervised). In this way there is no interruption in reaching the result but the user can review the translation if the result is poor.

4.3.3 Lesson Learnt: The Importance of Replication and the Fallacy of Measures

By keeping the same evaluation set-up (i.e. same document collection, same queries) we were able to compare if the interaction improved with the new design. The focussing of the query when seeing ambiguous translations recorded in the previous evaluation nearly disappeared. Thus the new layout did not stimulate a potentially negative behaviour and search effectiveness was not hampered.

The set of relevant documents selected by participants was analysed with respect to correct and incorrect relevance judgement. The number of non relevant documents selected as relevant was surprisingly high as participants were fluent speakers in the target language. Only 50% of the documents "wrongly" selected as relevant by participants had been actually

assessed; the other half was not in the CLEF pool and therefore had never been seen by any assessor[8]. The assumption that those documents were not relevant was then likely to be false.

The fact that documents which were rated as relevant by the assessors were not selected by the users was unexpected. Participants' behaviour was then analysed: they did not open (thus did not get the chance to read) the documents retrieved and judged the relevance from the title therefore discarding documents listed as relevant but that did not contained keywords in the title. This further confirms what was discussed by Mizzaro (1997) about the surrogate-based relevance judgement with the title surrogate performing worst. The use of relevance judgement as a way to measure the effectiveness of the whole system is therefore put into question.

4.4 Final System Assessment: Stepping Out of the Lab

4.4.1 Goal and Set-up

The evaluation of the final prototype was a summative evaluation of the work done over the three years (details in Levin & Petrelli 2004). In addition to the tasks and questionnaires previously used, participants were invited to bring a topic of their choice. The intent was to relax the controlled condition and move toward a more ecological setting, i.e. users with their individual task searching in their natural setting. Indeed the setting of this evaluation differed from the previous formative ones as: it was done at the user's premises and a single system was tested in all its aspects. Participants were eleven information professionals belonging to the user classes identified in the first user requirement study. They were business analysts, journalists, librarians and translators, people likely to use CLIR technology in the future. As the first showed (4.1 above) each class had different goals, expertise and attitude and we were interested in investigating how that variety of users experienced the Clarity system.

Users were assigned three tasks to test the different parts of the system (i.e. one-to-one cross language retrieval, multilingual retrieval, concept hierarchy browsing) and were required to bring a topic of their choice to test cross language retrieval from English to Latvian. The three tasks assigned were those used in the previous evaluations: knowing tasks and system performance in advance allowed us to outline expected users' behaviour and to question them if they did not conform.

Topics chosen by participants for the fourth task included: the Eurovision Song Contest, the restoration of Riga's Opera House, the status of Russians in Latvia, and Latvian foreign policy. Non of the participants could understand Latvian, all thought the system had retrieved documents relevant to their query and felt the translated titles and translations of terms found in the documents were helpful enough to be able to judge whether a document was relevant or not.

4.4.2 Data and Results

The data collected was both objective and subjective, though this time subjective measures (e.g. user's feeling and comfort) were considered more important than objective ones (e.g. completion time, documents retrieved). Results showed that the final system was robust, fast, accurate, easy and appealing to casual users. Comments were extremely positive and critiques were limited to very minor problems, e.g. keeping the translation selection from one turn to the next.

In the previous tests we discovered participants did not read the documents but judged the relevance from the title. This time we could ask the users' "why". Participants stated that "this is not real life! In real life I would read through even 200 documents if it matters" and "I have already got enough documents so I do not bother about others" and "I first collect what seems useful and read the material at a second stage". Each answer explains the behaviour of not reading the documents and judging the relevance from the surrogate, but it does not help in making relevance judgement a more reliable measure. The paradox discussed in 2.1. is emerging again: a more realistic evaluation with real task and use prevent comparison; an experimental setup for comparison is not ecological. Researchers should consider this tension when deciding to use relevance judgement as primary measure of IIR.

4.4.3 Lesson Learnt: User-Evaluation to Explore the Future

The "hybrid" evaluation adopted here, partially controlled and partially free, allowed monitoring the system improvement in respect to the previous prototype (the controlled condition) but also supported a shallow investigation on how the Clarity system could affect users' work. The fact that participants were required to formulate their own topic to search documents written in an unknown language prompted discussion on personal work (all the topics reflected their own work interest).

From the perspective of system development tasks variation becomes an issue when the major questions on design have been answered (Whittaker et al. 2000). Thus it is toward the end of the project that several tasks with different degrees of

[8] The reliability of the relevance assessment with a pooling method depends on the number and variety of search engines involved: the more and diverse the engines the more reliable the judgement.

complexity are worth testing. Asking the user to use their own topic to search the collection of Latvian newspaper articles allowed us to test the system in a broader sense and gave us the chance of discussing the condition of searching unknown languages (none of the participants could understand Latvian).

5. User-Centred Evaluation as an Inspection Method

Controlled user evaluations were instrumental to make the Clarity design progress (section 4.3). The setting adopted allowed to inspect both the software components (query formulation support, retrieval, and result visualization) and the user-interaction thus providing the Clarity team with a full understanding. The setting adopted is the topic of this section.

5.1. Separating and Inspecting IIR Components

The quality of retrieval is commonly measured at the end of the interaction. A single measure of the whole interaction (being that P&R or more user-centred ones, like in Su 1992) cannot expose what is really going on beside the interface. Experimenters would gain no insight into what were the negative factors: Was it the query formulation step that failed? Was it the search technique that was not robust enough? Was it the inappropriate display of the results? To gain full understanding, the different components of an IIR system has to be separated, measurements need to be taken during the whole interaction and related to the appropriate sub-task:

1. The user formulates a query in some form compatible with the modality required by the system input;
2. The system searches using internal algorithms;
3. The result of the search is presented to the user to evaluate.

Steps 1 and 3 pertain to the user interface, while step 2 is concerned with more technical issues (i.e. software architecture, searching algorithms). Inspecting each component separately allows locating where the problem is and why it occurs. This can be achieved by introducing intermediate measurement points. Figure 2 exemplifies this concept; the evaluation should determine: i) if the system adequately supports the user in expressing their needs; ii) if the search mechanism is effective given the user query; and iii) if the presentation of the results is good enough for the user to detect the relevant documents. Data should then be collected at each point.

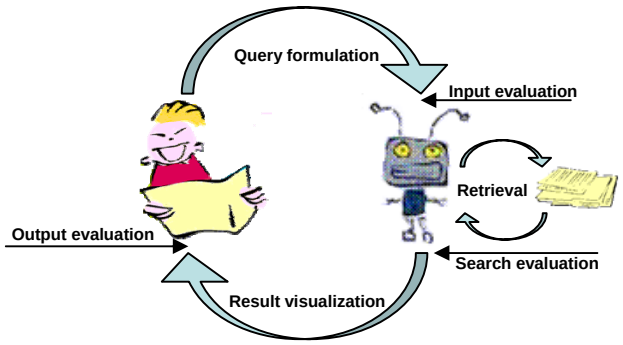

Figure 2. Measurement points in the IIR cycle.

The data collected should be both objective and subjective and should be analysed quantitatively and qualitatively. Indeed the aim of the evaluation is not just to assess the system performance, but to monitor the interaction and highlight cognitive aspects to be understood. The Clarity case showed that a rich set of data could be analysed in multiple ways; data collected included:

- *automatic log* (at each point): collect objective data (e.g. completion time, query terms, query number, clicks, document retrieved, documents seen) to be analysed quantitatively (to determine efficiency and effectiveness), as well as qualitatively (to find out problematic queries or unexpected behaviours);
- *observations of interaction*: collect information the log generally misses (e.g. user's attention, scrolling activity) to be analysed qualitatively;
- *interviews*: collect users' opinions and allow to ask questions about the observed behaviours (qualitative);
- *questionnaires*: collect users view on specific points and measure "how much" (e.g. system speed, reliability, layout) and therefore provide subjective data to be analysed in a quantitative way.

Quantitative analysis should measure effectiveness, efficiency and user satisfaction (as in the ISO definition of usability (van Welie et al. 1999) as inconsistency between objective and subjective measures is not unusual when complex systems are evaluated (Sellen 1995, Whittaker et al. 1993, Frøkjær et al. 2000, He et al. 2002). In Clarity (4.3) the most satisfying system (subjective) was not the best performer (objective). The qualitative analysis of users' behaviour provided an explanation for the phenomena and supported the designers in taking an informed decision on the final layout and interaction. Qualitative inspection should not be limited to the user-system interaction: search failures and mistranslation of proper names were identified through an analytical study of the recorded log (4.2.2). Qualitative analysis can also produce new research questions (e.g. why users did not open relevant documents 4.3.3) that can be explored in following evaluations.

The Clarity project shows that both quantitative and qualitative analysis should be done as they provide different insight; the first addresses the *what* the second the *why*. They should be considered as a kind of triangulation inside the same evaluation that allows covering multiple aspects and supporting cross-checking of the results thus improving the robustness of the findings.

5.3.1. Query formulation

Query formulation often determines the success of the whole interaction. CLIR is the ideal setting to show its importance, as the query translation is part of the query formulation step. The failure to correctly translate proper names was identified at this point by inspecting the log and comparing the query as entered by the user and the one used to search. Similarly, another CLIR user-centred evaluation (Petrelli & Clough 2005) revealed a numbers of improper translations of polysemous words that negatively affected the search.

However, query formulation is not just a matter of internal mechanism, the user interface also has a strong impact. Deciding which indicators are the most appropriate is crucial. Indeed, it could be more sensible to measure the success of the query formulation at the search evaluation point in terms of relevant documents retrieved: the higher the number of relevant documents retrieved the more effective the query session has been. Thus, precision can be used as a measure of effectiveness of the query formulation step.

To fully understand how much effort the user had to make to get a satisfactory outcome, efficiency needs to be addressed: the number of queries issued, the number of different terms used, and the average length of the queries can be considered as indicators of the level of user's effort (Belkin et al. 2003). The correct measure is then a balance between the effort spent in formulating a good query (efficiency) with its success (effectiveness). In Clarity (4.3.2) more effort was required by the supervised mode in terms of the number of actions, but the number of queries was less and the number of documents retrieved was higher.

5.3.2. Retrieval

The time the system spends in searching is a good indicator of efficiency, whereas the classical indices of precision and recall are both measures of effectiveness (Dunlop 2000). However, the examination of the performance of the search module should not be limited to the quantitative level; a qualitative analysis can be much more powerful in showing how the system is performing with respect to the user's query. For example, in the first Clarity evaluation (4.2), the queries "gene DNA disease" and "DNA genetic disorder", although semantically equivalent, returned different relevant documents. Qualitative diagnostic analysis is instrumental in detecting weak points and can inform us about how to improve search functionality. In Clarity, for example, discovering the wrong translation of proper names suggested a mechanism to by-pass that step. Examples of critical queries extracted from log data can be collected to build a corpus of actual user queries. This can be later used as user-simulated input during system tests without directly involving users (as done by Zhang et al. 2005).

5.3.3. Result Visualization

An interface that fails to display results effectively may hamper the successful use of the whole IIR system. Relevant documents retrieved can be used only if the user is able to identify them. An effective layout is important in general but becomes crucial when new forms of information retrieval are under investigation, e.g. multimedia IR, ubiquitous and mobile IR.

The effectiveness of the display strategy can be measured by considering the portion of relevant documents correctly identified out of the set of the documents retrieved, a sort of precision calculated on the basis of the retrieved set and not in relation to the whole document collection. This measure can be complemented with the measure of the portion of relevant documents wrongly judged by the user as non-relevant, i.e. the fallouts (Tague-Sutcliffe 1992).

Even though relevance judgements have been the core measure for a long time (Mizzaro 1997), they can be a weak measure if it is not known which documents have been judged by the assessors or how the assessment was done (on title, surrogate or

the whole document). This is likely to happen when a pooling system is used, as in TREC (Harman 1995), as only a subset of the text collection is assessed. Thus it is possible that other relevant documents are in the collection but are considered non relevant because no assessor has read them through (as we discovered in 4.3.3). Researchers should be aware of the potential problem and use this measure with caution, by for example introducing a further measure as an external judgement of the quality of the answer/solution given by the user (Hertzum & Frøkjær 1996).

6. Conclusions

The importance of user evaluation of IIR systems is becoming more widely accepted and typically involves user testing at the end of a project. Although it is essential to assess the system as a whole, user-centred evaluation also allows exploration of new forms of information access when performed iteratively. Evaluations are set at specific times in the project lifecycle to elicit different information: preliminary ideas can be tested using paper mock-ups and/or already existing systems; early tests with partial prototypes allow the system's potentials and limits to be explored; empirical evidence for design choices can emerge by evaluating consolidated prototypes and finding out how the system would perform in real life.

The complex interaction between a user and an interactive information retrieval system should consider the search engine as separated from the input mechanism and the result visualization. Only by measuring each subtask separately from the others is it possible to gain the micro-view needed to assess the effectiveness of each component.

The process of iterative evaluation, instead, provides the macro-view of IIR. Iterative evaluations can be done involving a limited number of participants, though the total number of people involved is comparable to a solid user-centred evaluation: in total 43 people participated in the Clarity studies. The strength of this approach lies more in the exploration of several solutions than on the definitive result of a single experiment. By iterative testing it was possible not only to state that the supervised mode for cross language information retrieval was more effective but less preferred by users than the delegated one (as also He and colleagues found (2002)) but it allowed us to understand why and to reach a solution that was both better performing and most preferred. The user centred evaluation then becomes a research tool that the experimenter can bend to the research needs; it is a form of longitudinal study applied to the IIR system.

User-centred evaluations are generally used to confirm or refute a hypothesis or to test a system. When used in formative evaluations as tool to explore innovative IR interactions, user-centred evaluation should aims at provide a complete picture; this can be achieved only if the range of data collected is rich in both objective (e.g. completion time, number of queries and terms) and subjective data (e.g. users' opinion), and if this range of data is analysed quantitatively as well as qualitatively. This makes possible a triangulation of data that better informs on the solution under study and does not stop at the "what" but reaches up for the "why".

Acknowledgements

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

Finally I thank the anonymous reviewers for their comments and suggestions that helped improve the organization and content of the paper, Christie Harrison for her help with the final version and Livio Milanesio for the drawings in figure 2.

References

Anick, P. (2003). Using Terminological Feedback for Web Search Refinement - A Log-based Study. In Callan, J., Hawking, D. & Smeaton, A. (Ed.) Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 2003, July 28–August 1, 2003, Toronto, Canada, 88-95.

Belkin, N., Cool, C., Kelly, D., Kim, JY., Lee, HJ., Muresan, G., Tamg, MC. & Yuan, XJ. (2003) Query Length in Interactive Information Retrieval. in Callan, J., Hawking, D. & Smeaton, A. (Ed.) Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 2003, July 28–August 1, 2003, Toronto, Canada,, 205-212.

Belkin, N., Cool, C., Kelly, D., Kim, G., Kim, Y-K., Lee, h-L, Muresan, G., Tang, M-C., Yuan, X-J. (2002) Rutgers Interactive Track at TREC 2002, in Proceedings of TREC 2002, Gaithersburg, November 2002. Available at http://trec.nist.gov/pubs/trec11/papers/rutgers.belkin.pdf [accessed 4.7.2006]

Benyon, D., Turner, P. & Turner, S. (2005) Designing Interactive Systems – People, Activities, Contexts, Technologies. Addison Wesley.

Borlund, P. (2000) Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation, 56 (1), 71-90.

Borlund, P. (2003) The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research, 8(3), paper no. 152 [available at http://informationr.net/ir/8-3/paper152.html]

Catani, M.B. & Biers, D.W. (1998) Usability evaluation and prototype fidelity: Users and usability professionals, Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 1331-1336.

Caulton, D.A. (2001) Relaxing the homogeneity assumption in usability testing. Behaviour & Information Technology, 20(1), 1-7.

Dumais, S., Cutrell, E. & Chen, H. (2001) Optimizing Search by Showing Results In Context. Proceedings of CHI 2001, 277-284.

Dumais, S., Cutrell, E., Cidix, JJ, Jancke, G., Sarin, R & Robbins, D. (2003) Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 72-79.

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

Draper, S.W. & Dunlop, M.D. (1997) New IR – New Evaluation: The impact of interaction and multimedia on information retrieval and its evaluation. The New Review of Hypermedia and Multimedia, vol. 3, 107-122.

Dunpol, M. (2000) Reflections on Mira: Interactive Evaluation in Information Retrieval. Journal of the American Society for Information Science. 51(14): 1269-1274.

Frøkjær, E., Hertzum, M. & Hornbaek, K. (2000) Measuring Usability: Are Effectiveness, Efficiency, and User Satisfaction Really Correlated? In proceedings of CHI 2000, 345-352.

Gould, J. D. & Lewis, C. (1985) Designing for Usability: Key Principles and What Designers Think. Communication of the ACM, March 1985, vol. 28, no. 3, 300-311.

Hackos, J-A T. & Redish, J. C. (1998) User and Task Analysis for Interface Design. Wiley.

Harman, D. (1995) Overview of the Second Text Retrieval Conference (TREC-2). Information Processing and Management, 31(3), 271-289.

He, D., Wang, J., Oard, D. & Nossal, M. (2002) Comparing User-assisted and Automatic Query Translation. Working notes CLEF 2002, 267-278.

Hertzum, M. & Frøkjær, E. (1996) Browsing and Querying in Online Documentation: A Study of User Interfaces and the Interaction Process. ACM Transactions on Computer-Human Interaction, 3 (2), 136-161.

Hounde, S. & Hill, C. (1997) What Do Prototypes Prototype? In Helander, M. G., Landauer, T. K., Prabhu, P.V "Handbook of Human-Computer Interaction" second ed., 367-381, North Holland, Elsevier.

Hughes, A., Marchionini, G., Wildemuth, B. & Wilkins, T. (2003) Text or Pictures? An eyetracking study of how people view digital video surrogates. Proceedings of the 2nd International Conference on Image and Video Retrieval CIVR 2003Springer Verlag LNCS 2728, 271-280.

Ingwersen, P & Järvelin, K (2005) The turn: Integration of information seeking and retrieval in context. Springer.

Levin, S. & Petrelli, D. (2003) Clarity Deliverable D6-2 "Report on Effectiveness of User Feedback CLIR System" [available at http://www.dcs.shef.ac.uk/nlp/clarity/reports/d6-2.pdf accessed 25.3.2006]

Levin, S. & Petrelli, D. (2004) Clarity Deliverable D7-1 "Report on Effectiveness of Clarity System" [available at http://www.dcs.shef.ac.uk/nlp/clarity/reports/d7-1.pdf access 25.3.2006].

Johnson. R (2005) An empirical investigation of sources of application-specific computer-self-efficacy and mediators of the efficacy-performance relationship. International Journal of Human-Computer Studies, 62, 737-758.

Kelly, D. & Belkin, N. (2004) Display time as implicit feedback: Understanding task effects, in Järvelin, K., Allan, J., Bruza, P. & Sanderson M. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, Sheffield, UK, 377-384.

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

McDonald, S.& Tait, J. (2003) Search Strategies in Content-Based Image Retrieval. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 80-87.

Mizzaro, S. (1997) Relevance: The Whole History, Journal of the American Society for Information Science and Technology, 48(9), 810-932

Nielsen, J. (2000) Design Web Usability, New Riders.

Oard (1998) A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA), Philadelphia, PA, October, 1998. Available at http://www.glue.umd.edu/~oard/research.html [accessed 4.7.2006]

Oard, D, Gonzalo, J., Sanderson, M, Lopez-Ostenero, F & Wang, J. (2004) Interactive Cross-Language Document Selection. In Information Retrieval, vol. 7, 205-228.

Over, P. (2001) The TREC interactive track: an annotated bibliography. Information Processing and Management. 37, 369-381.

Ozmutlu, S., Spink, A. & Ozmutlu, H (2004) A day in the life of Web searching: an exploratory study. Information Processing and Management, 40(2), March 2004, 319-345.

Petrelli, D. & Beaulieu, M. (2002) Clarity Deliverable D4-1 "User Studies for Clarity Design Scenarios and Observation of Real Users and Uses for Shaping CLIR Systems (Report on effectiveness of Clarity System)" [available at http://www.dcs.shef.ac.uk/nlp/clarity/reports/d4-1.pdf accessed 25.3.2006]

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

Petrelli, D., Hansen, P., Beaulieu, M., Sanderson, M., Demetriou, G. & Herring, P. (2004) Observing Users - Designing Clarity: A Case study on the user-centred design of a cross-language retrieval system. JASIST - Journal of the American Society for Information Science and Technology (JASIST) special topic on "Document Search Interface Design for Large-scale Collections", 55(10), 923-934.

Petrelli, D. & Clough, P. (2005) Using Concept Hierarchies in Text-Based Image Retrieval: A User Evaluation. CLEF evaluation forum.

Petrelli, D., Levin, S., Beaulieu, M. & Sanderson, M. (2006) Which User Interaction for Cross-Language Information Retrieval? Design Issues and Reflections. in JASIST Journal of the American Society for Information Science and Technology - special issue on "Multilingual Information Systems". 57(5), 709-722.

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

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

Robertson, S.E. & Hancock-Beaulieu, M.M (1992) On the evaluation of IR systems. Information Processing and Management, 28(4), 457-466.

Saracevic, T. (1995) Evaluation of Evaluation in Information Retrieval. Proc. SIGIR'95, 138-146.

Saracevic, T., Kantor, P., Chamis, A. & Trivison, D. (1988) A Study of Information Seeking and Retrieving. I - Background and Methodology. JASIS, 39(3), 161-176.

Sav, S., Jones, G., Lee, H., O'Connor, N. & Smeaton A. (2006) Interactive Experiments in Object-Based Retrieval. Proceedings of Image and Video Retrieval: 5th International Conference, CIVR 2006, Springer LNCS, 1-10.

Schusteritsch, R., Rao, S., & Rodden, K. (2005) "Mobile Search with Text Messages: Designing the User Experience for Google SMS" CHI 2005, April 2–7, Portland, Oregon, USA. 1777-1780

Sefelin, R., Tscheligi, M., and Giller, V., (2003) Paper prototyping – what is it good for? A comparison of paper-and computer-based prototyping, Proceedings of CHI 2003, 778-779.

Sellen, A. (1995) Remote conversations: the effects of mediating talk with technology, Human-Computer Interaction, 10, 401-444 .

Su, L. (1992) Evaluation measures for interactive information retrieval. Information Processing and Management. 28(4), 503-516.

Tague-Sutcliffe, J. (1992) The Pragmatics of Information Retrieval Experimentation, Revised. Information Processing and Management, 28 (4), 467-490.

Talja, S. (2002) Information sharing in academic communities: types and levels of collaboration in information seeking and use. The New Review of Information Behaviour Research - Studies of Information Seeking in Context, vol.3, 143-156.

Van Welie, M., van der Veer, G. C., Eliens, A. (1999) Breaking down Usability. Proc. INTERACT99, 613-620.

Virzi, R.A., Karis, D. & Sokolov, J.L. (1996), Usability problem identification using both low- and high-fidelity prototypes, Proceedings of CHI'96, 236-243.

Whittaker, S., Geelhoed, E. & Robinson, E. (1993) Shared workspaces: how do they work and when are they useful? International Journal of Man-Machine Studies, 39, 813-841.

Whittaker, S., Terveen, L. & Nardi, B. (2000) "Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI" Human-Computer Interaction, 15(2-3), 75-106.

Zhang, C., Chai, J., Jin, R. (2005) User term feedback in interactive text-based image retrieval. Proc. ACM SIGIR 2005, 51-58.