# Is Query Translation a Distinct Task from Search?

**Daniela Petrelli**
**Micheline Beaulieu**
**Mark Sanderson**
Information Studies

**George Demetriou**
**Patrick Herring**
Computer Science

University of Sheffield
Regent Court - 211 Portobello Street
S10 4DP – Sheffield – UK
{d.petrelli|m.beaulieu|m.sanderson}@shef.ac.uk
{g.demetriou|p.herring}@dcs.shef.ac.uk

## Introduction

The University of Sheffield participated in iCLEF 2002 using, as a test-bed, the prototype under development in the Clarity project. Clarity is an EU funded project aimed at developing a system for cross-language information retrieval for so-called *low density* languages, those with few translation resources. Currently translation between English and Finnish is supported; soon Swedish will be added and in the near future Latvian and Lithuanian.

Clarity is being developed in a user-centred way with user involvement from the beginning. The design of the first user interface was based on current best practise, particular attention was paid to empirical evidence for a specific design choice. Six paper-based interface mock-ups representing important points in the cross-language search task were generated and presented for user assessment as a part of an extensive user study. The study (reported in Petrelli et al. 2002) was conducted to understand users and uses of cross-language information retrieval systems. Many different techniques were applied: contextual enquiry, interviews, questionnaires, informal evaluation of existing cross-language technology, and participatory design sessions with the interface mock-ups mentioned above. As a result, a user class profile was sketched and a long list of user requirements was compiled. As a follow-up, a redesign session took place and the new system was designed for users who

- know the language(s) they are searching (polyglots);
- search for writing (journalists, translators business analysts);
- have limited searching skills;
- know the topic in advance or will learn/read on it while searching;
- use many languages in the same search session and often swap between them.

New system features were listed as important and the user interface was redesigned. Considering the result of the study the new interface allowed the user to dynamically change the language setting from query to query, hid the query translation and showed the retrieved set as ranked list primary.

Despite the fact that this new design was considered to be more effective, a comparison between the first layout based on the relevant literature and the new one based on the user study was considered an important research question. In particular, the choice of hiding the query translation was considered an important design decision, against the common agreement to allow and support the user in controlling the system actions. Thus the participation of Sheffield in iCLEF was organized around the idea of checking if the user should validate the query translation before the search is run or instead if the system should perform the translation and search in a single step without any user's supervision.

## Should the user check the query translation first?

The user interface is the means for the user to control the IR system: decisions and supervision has to be left to humans. This statement seems also well supported by empirical evidence in query expansion (Koenemann & Belkin 1996), and relevance feedback (Beaulieu & Jones 1998).

In CLIR, researchers assume that the user will type in the query in his/her own language, the system translates it to the document language(s) and performs the retrieval (Oard 1997). Controlling a CLIR system means that the user checks the query translation first and searches next as separated tasks. In the few fully implemented CLIR systems (Arctos and Mulinex (Capstick et al. 2000)) the control becomes the monitoring and refining of the query translation. The user's main job is to disambiguate translations coming from words with multiple senses.

Following this well traced path, the first Clarity interface design was based on the idea of give the user control of the translation before the search is done. A transparent user interface would show how the system translates the query terms and would allow users to modify, update and correct the translation before the search is actually performed.

When submitted to the users' judgement during the user study this interface was often criticised. Observers discovered that the full control over the system was not a requirement for most of the observed/interviewed users. Users seemed to care about the system's internal mechanisms only when something goes wrong and the result is not the one expected. Only in those situations users seemed to be interested in discovering what the system did and to eventually correct its behaviour (or modify their query).

A new interface was drawn for Clarity and the redesign completely contradicted the first solution: query translation and search were collapsed in a single step. This decision was supported by the observation that if a CLIR search engine works fine, it is able to disambiguate the query by itself and to retrieve mainly the relevant documents. In the unlucky case of a single word query with multiple senses the user can identify the problem by browsing the result[1] and reformulate the query exactly as it happens in all current search engines.

# The System Test-bed

## *The Architecture*

Although iCLEF did not require to have a running CLIR system behind the interface, the University of Sheffield decided to conduct the test under conditions that resembled a "natural setting" as much as possible. Thus the whole Clarity prototype was used, not only the user interface. The system is physically distributed with the user interface and few correlated services in the UK (at the University of Sheffield) and the cross-language retrieving core in Finland (query translation and searching performed at the University of Tampere). SOAP services facilitated communication between the two sites. The tested prototype can translate queries from English into Finnish and searche documents in both English and Finnish at the same time. For iCLEF, the search was limited to only Finnish documents, thus the user typed a query in English and retrieved documents in Finnish.

## *The User Interface*

As discussed above, the Clarity user interface was designed for a target user class of people who know the languages they are searching. This implies that mechanisms like document or title translation are not needed. One of the iCLEF requirements was, however, to recruit users who do not know the language they are searching. Thus the user interface was slightly modified to support this type of user. The first change was to include under the original title in Finnish, a translation in English. The translation used a word-by-word mechanism; multiple translations for a single word were displayed in sequence separated by commas. In our standard interface no document or summary translation was offered to the users, for iCLEF, however, some other feedback on the document content had to be provided in English. It was decided to list extracted document keywords in Finnish with a corresponding translated set and a list of proper names. Both these facilities were provided on the retrieved documents. Despite the fact that the mechanisms used were sometimes very weak (e.g. effectiveness of keywords extraction strongly depended on the relevance of the retrieved documents) it was considered important to test the two features to get an idea on their effectiveness.

[1] As discussed above Clarity target users are polyglots searching the languages they know.

# Planning the Experiment

## *Hypothesis*

As discussed above, the result of the user study pushed Clarity designers to revise the first interface. The new user interface hid the translation mechanisms in favour of a simpler layout and interaction. The CLIR engine was trusted to be robust enough to support a "delegation" from the user to the system more than a "supervision" of the user over the system.

This choice was well motivated by the result of the study, but it has not been proved that this is what occurs in the reality. It might be that the cross-language technology is not yet robust enough for supporting this collaboration or that it works only for users who know the target language. Indeed the fact that the user knows the target language might deeply affect the retrieving. Users who do not know the language they are searching and do not use the retrieved information might feel more confident on the process if they can understand a minimum of the CLIR process (i.e. if they control and supervise the query translation). The possibility of controlling the query translation may increase the confidence that the set of information retrieved is reasonably complete. For such users, more control over the system might be desired.

## *Experimental Conditions*

Our research goal was to test if the query translation step had to be considered as a separate task with respect to searching or if it was a better design solution to hide it. If our iCLEF subjects performed better on the interface that does not display query translation (referred to as NoTrans) than on the interface that forces a query translation check before the search (referred to as Trans), then hiding the translation would be considered the best design choice in general.

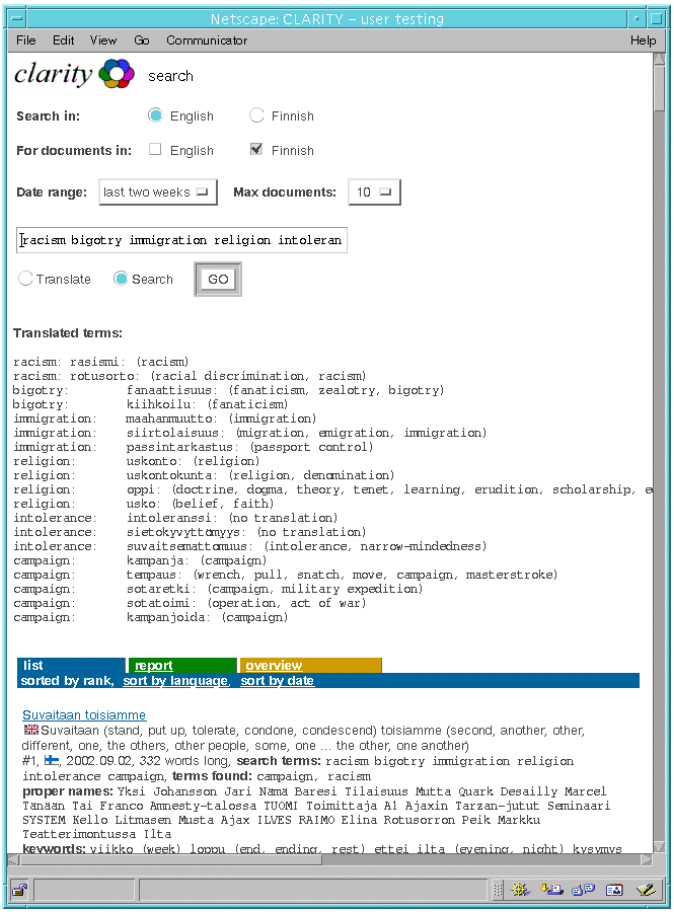


Figure 1. The layout that forces the user to check the translation (Trans).

The layout of the two interfaces corresponding to the conditions was kept as similar as possible to avoid the interface design affecting the interaction. The only difference was in the control the user had on the query-translation process. In the Trans design (Figure 1) the user received first a full feedback on how the system translated the query from English into Finnish. The user could then change the English query and again review the Finnish translation (note that back translation into English was displayed in brackets so that the user saw alternative meanings for the same word). When the users were satisfied with the query they clicked for searching.

In the NoTrans design (Figure 2), the translation was not displayed as default and users ran a search as with a standard IR system. To reduce the differences between the two layouts, a button to display the query translation on top of the ranked list was added. When clicked, the button offered the user feedback on what the system did; however translation and search were not separated.

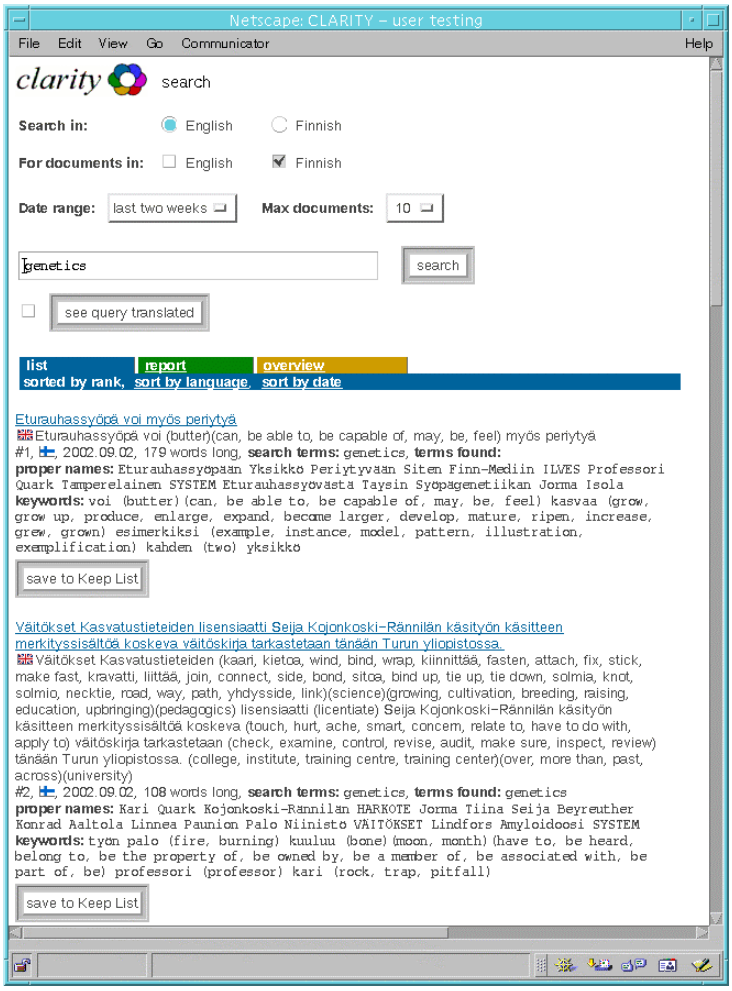


Figure 2. The layout that hides the translation (NoTrans).

## *Experimental Design*

As required by iCLEF, a within subject design was adopted where each participant used both interfaces. The iCLEF matrix was used to counterbalance the learning effect. Objective and subjective measures were collected. A log recorded the queries submitted, the retrieved set, and the relevance judgement given by users. Questionnaires were used throughout the experiment to collect subjective data. All were designed to get the strength of agreement with a clear statement (a 5 point Likert scale), as for "I always understand why a document has been retrieved" with the scale from "strongly disagree" to "strongly agree". While the questionnaires filled after the tasks completion aimed at collecting opinions on topics and systems, the preliminary one was designed to gather information about user expertise with computer and search and to figure out their general attitude to IR. Whenever

possible, participants were videotaped or observed by the experimenter in order to collect qualitative information about the interaction.

Participant's performance in terms of number of documents selected as relevant was the main measure required in iCLEF. Users were asked to move interesting documents to a "keep list" and then to judge those as "somewhat relevant", "highly relevant" or "not relevant". Time was not considered due to the fact that accessing a remote service in Finland did not assure a constant time across users.

# Running the Experiment

## *Participants*

Six people (students and research associates) recruited at the departments of Information Studies and Computer Sciences participated in the experiment. They were paid £15 for their participation. Of those, only the four fitting the iCLEF matrix were considered here. All participants were English native speakers and none of them had any (passive or active) knowledge of Finnish. All considered themselves good or expert computer users (mean= 4.26, sd= 0.96) and fair searchers (mean= 3.75, sd= 0.50), search the web often (mean= 4.50, sd= 1), seldom the library (mean= 3.75, sd= 0.50), and rarely other DB (mean= 2.25, sd= 1.26).

Participant attitude toward searching was collected in nine questions answered before starting the interaction. In general these people seems to be positive with respect to the search task: they were confident they would find what they are looking for and be able to reformulate and feel in control of the search process. Opinions were less homogeneous when addressing the need for help to formulate a query or the understanding of the reason for retrieving a certain document. Also the perception of the "advanced" or "Boolean" search was not straightforward: it seemed those were considered powerful tools but rarely used.

## *Procedure*

Participants used the computers available in the student laboratory and used the system in parallel. Indeed this was later discovered to be a possible reason for bad performances since the system was not designed for multiple simultaneous accesses.

At arrival, users received a page describing the experiment and its purpose. They ran a training query with both systems and then conducted two searched on each system. Users were provided with a context-based scenario that described the task (Borlund & Ingwersen 2000). After each search, participants filled in a questionnaire about the topic; familiarity, difficulties in query formulation or judgement, and confidence were the questions addressed. Opinions about the system were collected after the two tasks were completed, a further questionnaire comparing the two systems was filled at the end.

# Results and Discussion

Results returned by iCLEF are shown in the table below. The effectiveness of users was determined using Van Rijsbergen's *F* measure, where user judgements were compared to those made previously by assessors.

| **System** | **F** |
|---|---|
| NoTrans | 0.081 |
| Tran | 0.061 |

Table 1 Official iCLEF results.

The low numbers were due to the fact that of the 16 tasks only in four cases participants were able to retrieve relevant documents. Of the four, three were conducted on the NoTrans system, one with the Trans interface. On the topic side the four successful searches were on only two topics: there was no result for the other two. Finally, one participant was successful in two tasks, two were successful in one, and one participant failed to find any relevant document in all tasks.

On closer inspection, it seems that the users performed slightly better than it first appeared. Documents marked as "somewhat relevant" were ignored when calculating *F*. Thus, for one user search, instead of the 10 documents marked only six have been used. A recalculation including somewhat relevant will be conducted in the future.

Although sometimes a few relevant documents were retrieved and the participant failed in identify them, more commonly the relevant documents were not retrieved at all. Such a bad performance was explained considering some technical aspects of the system used. Participants accessed the system in parallel, but the system was not designed to support multiple accesses. Later it was discovered that simultaneous accesses would write results to a common file. Right now we cannot say how much this misbehaviour affected the data collected.

The CLIR engine was also not as efficient as expected. Indeed the users were not able to formulate queries that retrieved all the relevant documents. This seemed to be related to the fact that all the possible translations were used. Thus for example for the word "green" the less common senses "vegetable" and "unripe" were translated in Finnish. A deeper analysis of the queries used will clarify this point.

Finally the bad performance might be related to participant personality. The user who failed all the tasks was observed typing "bobby sand" while searching for "hunger strikes" (Sands was a well known hunger striker). The system translated this query wrongly searching for "policeman" and "beach". This participant was very happy about trapping the system.

Respect to the user opinions a full analysis of the questionnaires is not very useful given the poor performance. However it is worth noting that none of the participants mentioned the Trans system as easy to learn or use, or as the most preferred. Comments on the Trans interface generally complained about the translations into figurative or metaphoric meanings. However the biggest complaint was on the system slowness: sometimes users had to wait minutes before seeing the retrieved set.

# Lessons Learnt

First we learnt that using a real system could be very risky but also extremely rewarding in terms of usability test. The technical problems we had hampered the collection of a reliable set of data, but being in the worst situation pushed the participants to try as many paths as they could. So we observed participants try to use "+" and "-" to disambiguate or restrict queries as well as proper names and acronyms. The system did not support Booleans nor proper names, but now we know how important they are.

The measurements did not clarify the initial question on which design is the best one, however, by observing participants we could imagine the cognitive mechanisms behind the layouts. Almost all the observed people were puzzled when the Trans interface showed the query translation; all users expected to see the result of the search so patiently waited for a while. After the translation was shown all the observed people modified the query before searching. Being more specific without having seen any documents might prevent the retrieving of relevant ones when those contain mainly the generic terms.

When facing ambiguous terms, participants tried to disambiguate the query by using the Finnish word. This made the system perform worse and the user become frustrated. This could probably be avoided by offering a layout that allows users to dynamically select the right sense. However this solution does not guarantee a better performance.

# Acknowledgements

Our gratitude 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 experiment. We thank also the European Commission for funding Clarity (contract no.:IST-2000-25310).

# References

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

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

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

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

(Petrelli et al 2002) Petrelli, D., Hansen, P., Beaulieu, M. and Sanderson M. User Requirement Elicitation for Cross-Language Information Retrieval. To be published in the Proceedings of International Conference on Information Seeking and Retrieval in Context – ISIC 2002, Lisbon, September 2002.

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