SICS Technical Report T2009:03
ISRN: SICS-T–2009/03-SE
ISSN: 1100-3154

Texts and Language – Interactivity and Context

Experiences of Interactive Cross-Language Experiments
at SICS 2003-2005

Preben Hansen and Jussi Karlgren
{preben,jussi@sics.se}

2009

Swedish Institute of Computer Science
Box 1263, S-164 29 KISTA, SWEDEN

Abstract

This technical report collects three years of experimentation in interactive cross-language information retrieval by SICS in the annual Cross-language Evaluation Forum (CLEF) evaluation campaigns 2003, 2004, and 2005. We varied simulated task context and measured user performance in document assessment task to find that choice of language and task context indeed have effects on the amount of efforts users need to expend to achieve task completion.

Keywords: Context, Simulated Domain and Work-Task Scenario (SDWS) methodology, Cross-lingual information access, Information Retrieval, Relevance assessment, Reading.

1 Cross-lingual information access as a research field and as an application

Cross-lingual information access is an application field in its own right. There are several use cases for technology, which retrieves relevant documents from a multilingual collection without requiring specification in every target language by the user.

There are obvious reasons for research to develop the various technologies necessary for building a cross-lingual information access service, but the field of cross-lingual information access is also concerned with formulating requirements for the same technology, directions for research, and schemes for applying the resulting technology and algorithms to real world tasks.

Cross-lingual information access as a user task is more complex than monolingual information access, and study in aspects of what challenges users experience in cross-lingual relevance assessment may shed light on factors of information access that are general and possible to apply profitably in the monolingual case as well.

As all information access research, study in cross-lingual information access is evaluation oriented. This present collection of studies has been performed at some of the annual Cross-language Evaluation Forum (CLEF) information access evaluation campaigns where various technologies have been applied to shared tasks.

1.2 CLEF and iCLEF

Since 1992 the annual Text Retrieval Conference (TREC)[1] organized by the National Institute of Standards and Technology in the United States, the most important forum for evaluating information retrieval system performance, has organized a cross-lingual retrieval evaluation track and an interactive retrieval track that evaluates various aspects of human operation of information retrieval systems.

In the year 2000 the European Commission in cooperation with TREC initiated and organized the yearly Cross-Language Evaluation Forum (CLEF)[2], which has gathered research groups interested in experimenting with European languages. Since year 1998 the Japanese evaluation project NII-NACSIS Test Collection for IR Systems (NTCIR)[3] has semi-annually evaluated task-oriented cross-lingual performance of retrieval systems.

[1] http://trec.nist.gov/

[2] http://clef.iei.pi.cnr.it:2002/

[3] http://research.nii.ac.jp/ntcir/

All of these evaluation platforms have a common empirical ground, in that they distribute given topics to participants who then submit results based on their system performance for judging by the organizers. In the following experiments we have used test queries, test collections, and relevance assessments established for them in the interactive track of CLEF - iCLEF[4]. The test protocol for that year's interactive experiments are given in detail in an overview paper (Gonzalo and Oard, 2003).

The aim of CLEF is to support the development of research communities devoted to cross-language research by facilitating cooperation between research groups with common interests, and by supporting empirical evaluation of tools and algorithms for information retrieval systems operating on European languages. Furthermore, CLEF aids the creation of "test-suites of reusable data which can be employed by system developers for benchmarking purposes." There are strong established links with the two other initiatives mentioned above. As stated on CLEF webpage: "The final goal is to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness on the global marketplace." [5]

1.3 Purpose and goal with this report

This reports collects the experiments performed by SICS in the iCLEF evaluation track. All have been published in the working notes and the proceedings of CLEF, and a longer version has been published as a journal paper – this report is to collect the material in convenient form.

The papers presented and discussed below have previously been published. The first three papers have been presented at the Workshop of the Cross-Language Evaluation Forum between the years of 2002-2004. The 4th paper is a journal article published in 2005:

(1) Jussi Karlgren & Preben Hansen. (2003). Cross-Language Relevance Assessment and Task Context. In: Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck. (eds), Advances in Cross-Language Information Retrieval. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 2002. Lecture Notes in Computer Science 2785. Springer 2003, pp. 383-391.

(2) Jussi Karlgren & Preben Hansen. (2004). Continued Experiments on Cross-Language Relevance Assessment. In: Carol Peters, Martin Braschler, and Julio Gonzalo (eds.). Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Lecture Notes in Computer Science LNCS 3237. Part III. Interactive Cross-Language Retrieval, (pp. 468-470). Springer-Verlag: Heidelberg.

(3) Preben Hansen, Jussi Karlgren & Magnus Sahlgren. (2005). Bookmarking, Thesaurus, and Cooperation in Bilingual Question Answering.
In Working Notes for the 4th Workshop of the Cross Language Evaluation Forum, Bath, England, September 2004.

[4] http://nlp.uned.es/iCLEF/

[5] http://www.clef-campaign.org/

(4) Hansen, Preben & Karlgren, Jussi (2005). Effects of foreign language and task scenario on relevance assessment. Journal of Documentation, Vol.61 (5), 2005, pp. 623-639.

2 Summarization of approaches and results

In the following section we will give a short description of each of the papers given in the appendix of this report. Each paper will be described according to the following sub-sections:

a) A short *description*
b) *Research question(s)* addressed;
c) *Type of experiment* and *experiment set-up*
d) *results achieved*, followed by a
e) short *Discussion*

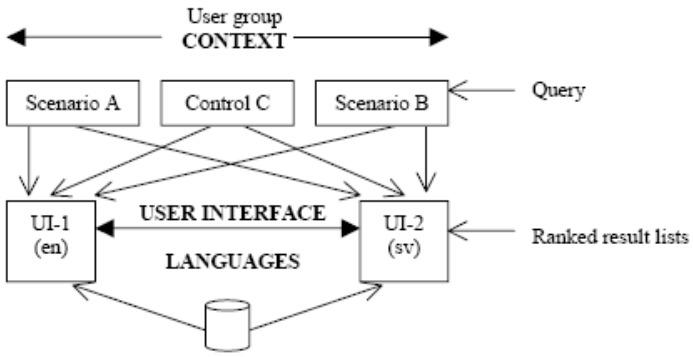
2.1 Paper #1: [2003] Karlgren, J. & Hansen, P. Cross-Language Relevance Assessment and Task Context

This is an experiment on how users assess relevance in a foreign language they know well is reported. Results show that relevance assessment in a foreign language takes more time and is prone to errors compared to assessment in the reader's first language. The results are related to task and context and an enhanced methodology for performing context-sensitive studies is reported.

Research questions addressed
In our first paper, our hypotheses were that results for a foreign language would be more time-consuming and less competent than those for the first language.

Experiment set-up
In this study, we want to relate the relevance assessment to a specific task situation, i.e. the subject will be given a semi-realistic situation including a domain description, and then we will investigate if the relevance assessment situation involves criteria beyond topicality. The following task-based experimental design were used (figure 1)

Figure 1. Task and scenario-based experiment design



In this study we defined and used a first version of a Simulated Domain and Work-Task Scenario (SDWS) methodology (Figure 2). See appendix A for a full example of a SDWS

General descriptions: Domain: Work-task description: Situational description: Topic: Search task description:

Table 1: The framework of the general description level as well as the situational description level.

The study involved 12 participants divided into 3 groups. Groups A and B were given a workplace scenario involving a domain with relevant work-tasks. Group C was given the i-CLEF queries without context information. Each scenario had 4 participants. English and Swedish were used. Four CLEF queries were used from the 2002 year's interactive track.

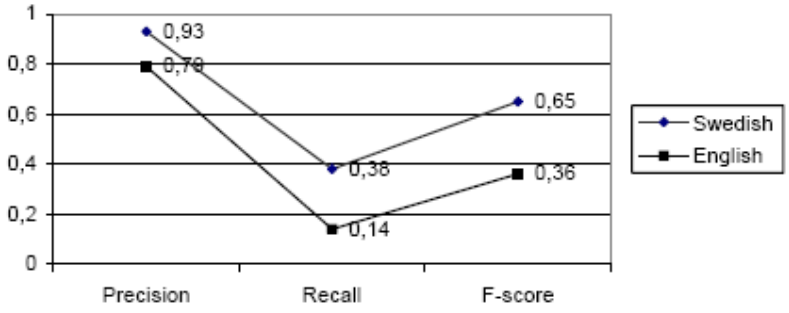
We used a sets of ranked result lists of length between one and two hundred lines were produced in Swedish using Siteseeker, a commercial web-based search system by Euroseek AB, on the TT CLEF corpus and English using Inquery on the LA Times CLEF corpus. The ranked lists were presented to the participants, varied by order and language in a simulated search interface. Four relevance categories were used: "not relevant" "somewhat relevant", "relevant", and "don't know".

Outcome

A. Foreign-language texts took longer to assess and were assessed less well.

Assessing texts in English (30 s average assessment time) took longer than for Swedish (19 s). Given the extra effort invested into reading the English texts it is somewhat surprising to find that the results of the assessments were significantly less reliable for English than for Swedish as well (figure 2).

Figure 2: Retrieval results

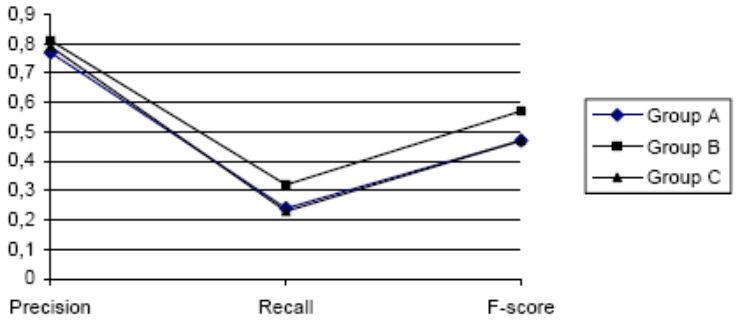


All differences between English and Swedish were significant by Mann Whitney U; p > 0,95. Assessments were judged by how well they correspond to the CLEF official assessments; precision and recall are calculated with respect to the known relevant documents found in the retrieved and presented set of documents. The average length of an English article is over seven hundred word, whereas the Swedish articles are of an average length of just over four hundred. The length difference could account for part of the assessment time difference, but since the length of the article correlates very weakly with assessment time (Spearman's Rho = 0.3) that explanation can be discounted.

B. Task focus may have an effect on assessment performance
No significant differences between scenarios (cf. Figure 3) could be found, other than a tendency for group B to perform better (p > 0,75; Mann Whitney U) than group A and C (the control group).

As found by questionnaire, group B invested less effort in topic and more in task related aspects of relevance than did group A, which may be a tentative explanation for the tendency; this relation needs to be investigated further before any conclusions can be drawn, however.

Figure 3: Retrieval results by task.



C. Methodology: Enhanced methodology for context-sensitive studies.
In order to perform relevance assessments in a specific work-task situation, we designed an enhanced contextual framework description – SDWS (Simulated Domain

and Work-Task Scenario). Basically, the SDWS has 2 main sections with 2 subsections each:

D. Relevance judgment aspects related to task
One very important finding was the fact that we assumed that aspects of the relevance judgment taken into account would extend beyond traditional topicality. We added two more levels related to our domain and task-based scenario approach. After each query, the participants were asked what aspects of relevance judgments were of any importance for their assessment. The following relevance aspects were used:

- Relevance judgement aspect related to the task domain
- Relevance judgement aspect related to the task given to the participant
- Relevance judgement aspect related to the topic of the query

Merged, the two groups (A and B) used the domain related aspect in 12% of the relevance judgement cases, the task related aspect in 46% of the cases, and the topic-related aspect in 42% of the cases. Notable is that 36% in the A-group and 61 % in the B-group marked that their assessments were related to task.

Another interesting observation is that nobody in the group B reported using the domain-related aspect in assessments. Group A had a level of 44% on topic-related aspect and 36% on task-related aspects.

Discussion
The results are quite convincing.

a) Time matters: Relevance assessment in a foreign language, even a familiar one, is more time-consuming and more difficult than in one's first language.
b) Tasks seem to matter: Generally, traditional information retrieval experiments are based on algorithmic and topical relevance. In this study we have seen that other aspects do count in the relevance assessment.
c) Methodology and Scenario: Furthermore, we have a weak but interesting indication that the Simulated Domain and Work-Task Scenario applied may have an effect on the assessment performance.

2.2 Paper #2: [2004] Karlgren, J. & Hansen, P. Continued Experiments on Cross-Language Relevance Assessment

This experiment was a smaller study than the previous and due to an unreliable logging functionality employed in the study, some of the data we needed to investigate did not surface. The experiment is on how users assess document usefulness for an information access task in their native language (Swedish) versus a language they have near-native competence in (English).

Research questions addressed
Examine what mechanisms might be introduced to the retrieval situation to close the gap between the two linguistic conditions. This year we focused on the user interface and investigate the utility of an interface detail, which invites the user to deliberate the selection of documents further.

Experiment set-up
8 participants were involved. Data: The Swedish TT Telegrambyrå CLEF corpus was used as well as the English LA Times corpus. The experiment used the Clarity system with an added interface functionality of a bookmarking tool. Searches were performed in Swedish and retrieved results in either Swedish/English. Ranked lists with news item headlines visible in original language. 4 CLEF queries were used.

The participants were asked to answer an initial questionnaire. They were then given the TREC topic description of four queries and were allowed to formulate queries freely in Swedish, to inspect the resulting list, to select documents for reading, to reformulate the query, to save or delete documents from the bookmark panel. Between each test query the participants were asked to answer a fixed set of questions related to the retrieval system.

Outcome and Discussions
Results show that relevance assessment in a foreign language takes more time and is prone to errors compared to assessment in the reader's first language.

1. Native (Swedish) documents were viewed for reading more often (9.5 %) of Swedish documents viewed) than foreign documents (7.4 %)[6]. This calls into question if the headline does have an effect.
2. Foreign (English) documents were discarded from the bookmark panel more often after having first being saved (40% or 35 out of 80 saved documents) compared to native documents (15% or 14 out of 80 saved documents)[7]. The question is then if the users were less confident in their first impressions of foreign language documents.
3. Searches in native (Swedish) documents were reformulated more often than searches in the foreign language (English). Recycling terminology from the target set seems to have an effect.

Human reading performance needs to be studied further. Do headlines have an effect? Do different levels of reader confidence in different languages have an effect?

2.3 Paper #3: [2005] Hansen, P., Karlgren, J., Sahlgren, M. Bookmarking, Thesaurus, and Cooperation in interactive Bilingual Question Answering

This exploratory study on information access behavior in a multi-lingual context presented involves several different contextual. This interactive CLEF experiment was designed to measure three parameters we expected would affect the performance of users in cross-lingual tasks in languages in which the users are less than fluent.

Research questions addressed

---

[6] Significant by x2;p >0.95

[7] Significant by x2;p >0.999

- How will topic-tailored *term expansion* on query formulation affect the performance of users in cross-lingual tasks?
- What is the effect of a *bookmark panel* on user confidence in the reported result
- How do people cooperate and *collaborate* with a partner during a search session performing a similar but non-identical search task.

Experiment set-up

The target language was French. No translation service was provided. All 16 queries were formulated in French, all documents were displayed in French, and all answers were given in French.

The subjects were primarily Swedish speakers and moderately competent in the target language. No fluent French readers were accepted. The 8 participants were paired together two-and-two. 75 per cent of the users were female and the average age of all users was 34. None of the users had any experience performing a similar experiment. Average experience with online searching was 7,5 years. 8 participants were paired together in sessions. The two participants were sitting at a table, opposite to each other, so that they could see each other's face-to-face. Each of them had a search terminal with network access and the search system installed. The table had enough space to write notes. Each participant also received a topic protocol including a set of questionnaires individually designed as to the matrix.

The questionnaires contained three sets of questions: one initial questionnaire, a questionnaire suited to each of the two systems tested (with and without query expansion) and a final questionnaire.

For the data collection, we observed the subjects when performing their search tasks. The observer had a copy of the set of queries the subjects were assigned and used a notebook to collect data for each specific query performed according to a set of pre-defined variables such as dialogues and conversations made for each query pair. For the analysis of the data, all the written notes from each session were coded and analyzed by content.

Three Simultaneous Experiments. The study presented involves several different contextual aspects and is the latest in a continuing series of exploratory experiments on information access behavior in a multi-lingual context. In this study, we measured three parameters that we expected would affect the performance of users in cross-lingual tasks in languages in which the users are less than fluent.

- Firstly, we measure the effect of topic-tailored term expansion on query formulation. Subjects were first given eight queries without term expansion capabilities and then eight with an added window where a French word could be entered to retrieve up to five suggestions of related terms. The thesaurus used for expansion was generated automatically from parallel corpora of EU legislation by GSDM methodology
- Secondly, introducing a new component in the interactive interface, we investigated - without measuring by using a control group - the effect of a bookmark panel on user confidence in the reported result. For this we used a bookmark panel. All users were given this feature in all queries. Users could mark an arbitrary selection from a displayed document and bookmark it to be used for answer extraction at the end of the task, as described below.

- Thirdly, we ran subjects pair-wise and allowed them to communicate verbally, to investigate how people may cooperate and collaborate with a partner during a search session performing a similar but non-identical search task. This is a novel and rather unexplored component of information access system evaluation when performing search tasks. The users were given support to perform their information access tasks in partial collaboration with other subjects. The subjects performed the search in pairs. The sequences of queries were kept different within pairs according to the i-CLEF experimental matrix: the subjects never worked on the same query simultaneously. Subject communication was logged by encoding communication in one of a limited set of categories such as "vocabulary question" "system operation question".

The text retrieval engine used for our experiments is based on a standard retrieval system being developed at SICS. The system allowed: *ranked list*, a *document display window*, *a bookmark list*, text selection window, bookmarked items can be *checked* (and un-checked again), and the highlighted text snippet is copied to the *answer window* and the document title is automatically copied to the *reference window*. The system also has a *thesaurus* component.

In this study, the *relevance* aspects are:

a) "Saved answer"
b) "Final Answer", and
c) "Uncertain Answer"

The first bookmarked selection is copied into the answer display field by default and the user has the option of manually editing or entering an answer into the answer display field. The system also has a *thesaurus* component.

Outcome and Discussions

In short, this study showed that:

A. Thesaurus and Term Expansion. The thesaurus was not useful due to its limited coverage, in spite of it having improved retrieval results in a wholly automatic setting. Subjects were frustrated by its inherent unpredictability and its patchy coverage. Most subjects tested it once or twice and did not use it thereafter.

B. Bookmark Panel. Users were happy about the bookmark panel and commented on it in a positive way.

C. Cooperation. We found that users actually did cooperate and collaborate. The subjects communicated around 4 categories:

- Topic
- Search strategies
- The language (Vocabularies and Translation)
- System functionalities

The data set from four sets of participants is too meager to draw any more fine-grained conclusions, but provisionally we were able to note that the more participants communicated, the more similar their results and answer turned out to be.

One of the lasting results will be the continued development of evaluation methodology. We need a more robust framework for studying collaboration in information access – a task, which is naturally cooperative rather than individual. This study points at one possible route to take: free form communication, interactive turns categorized by an experiment conductor, tasks similar but separate.

2.4 Paper #4: [2005] Hansen, Preben & Karlgren, Jussi. Effects of foreign language and task scenario on Relevance Assessment

Paper 4 is an extended version of the study reported in paper 1 including a more developed research setting and with additional results.

A controlled interactive information retrieval experiment was performed on how readers assess relevance of retrieved documents in a foreign language they know well. A two-level scenario description framework was applied to facilitate study of context effects on the assessment process.

Research questions addressed
How to assess relevance in a foreign language they know well compared with their native language? Hypotheses: assessments done in foreign language would be more time-consuming and less competent than for the first language

Experiment set-up
The relevance assessment was related to a specific task situation. The participants were given a more realistic task-situation that also included a domain description. This was done in order to investigate if the relevance assessment situation involved criteria beyond topicality.

Participants were grouped to participate in different *task scenarios* and the SDWS framework description (app. A) were used including two different two-layered domain scenarios. The scenarios were derived from two real-life work-task situations and designed as follows: *the first level* contains a short description of the *domain* and of general *work-tasks* or routines usually performed within this domain. *The second level* contains a situational description including the *topic* of the query and a *search task* description:

In this way, the scenarios would allow the participant
a) a broader understanding of the actual information-seeking situation, and
b) to perform a task-oriented interpretation of the relevance.

So, for each scenario there was one general description level and four different situational descriptions corresponding to the four i-CLEF queries selected for the study (C053; C056; C065; and C080). The designed domains for scenario A and B (see below), were assigned randomly to the participants:

Scenario A (Group A):	Domain:	Monitoring news and translation services
Scenario B (Group B):	Domain:	Information specialist /Consultant

Outcome and Discussions

The results showed that:

A: Relevance assessments

Relevance assessment takes longer in a foreign language than in a user's first language.

a) despite given the extra time in reading the foreign language, the quality of assessments by comparison with pre-assessed results is inferior to those made in the users' first language.

b) assessing texts in English (27 s average assessment time per document) took longer than for Swedish (20 s) ($p > 0.95$; Mann Whitney U) as shown in the figure below:

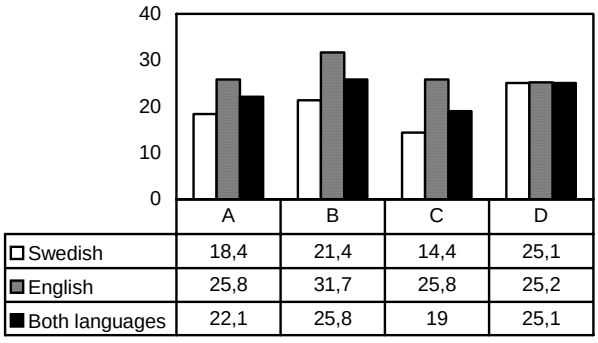
B. Scenario effects on assessment effort

The simulated scenarios did make a difference for the assessment process, both by

- measured access time and by
- self-report by subjects.

Scenario has quite significant effects on interaction time ($p > 0.995$; Kruskal-Wallis). This is shown in Figure 4 below. The table also shows that scenario has effect on time within the native language. A considerable difference in time can be observed between the group with scenario (A) and the group with a terse query, while in a non-native language, the interaction time was almost the same for each group except for the scenario B.

Figure 4. Average assessment time by scenario. [s]



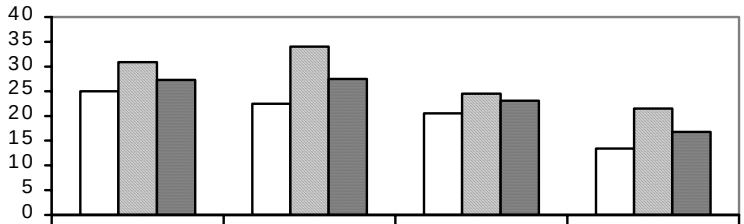
	A	B	C	D
Swedish	18,4	21,4	14,4	25,1
English	25,8	31,7	25,8	25,2
Both languages	22,1	25,8	19	25,1

C. Topic effects on assessment effort

No effects on results by traditional relevance ranking were detectable. This is a strong argument for extending the evaluation measure to cater for context effects in addition to the more traditional experimental topical relevance measures.

Topic has an effect on assessment time ($p > 0.95$; Kruskal-Wallis) as shown in Figure 5 and this is not surprising since some retrieval topics are simply more difficult in a given collection.

Figure 5. Assessment agreement across language and topic



	53	56	65	80
Swedish	25	22,5	20,5	13,4
English	30,9	34	24,5	21,5
Both languages	27,3	27,5	23,1	16,8

The discussion above can be summarized into a table for a more comprehensive overview of the experiments performed.

Table 2: Summary of the four studies

	Paper #1	Paper #2	Paper #3	Paper #4
Research questions	How to assess relevance in a foreign language they know well compared with their native language? Hypotheses: assessments done in foreign language would be more time-consuming and less competent than for the first language	Investigate the utility of a user interface functionality detail, which invites the user to deliberate the selection of documents further and if this had any effect language.	Three problems were addressed: How will topic-tailored term expansion on query formulation affect the performance of users in cross-lingual tasks? What is the effect of a bookmark panel on user confidence in the reported result How do people cooperate and collaborate with a partner during a search session performing a similar but non-identical search task.	How to assess relevance in a foreign language they know well compared with their native language? Hypotheses: assessments done in foreign language would be more time-consuming and less competent than for the first language
Experimental Set-up	Relate assessments to task situations through a model of a semi-realistic situation and domain description	The participants were asked to answer an initial questionnaire. They were then given the TREC topic description of four queries and were allowed to formulate queries freely in Swedish, to inspect the resulting list, to select documents for reading, to reformulate the query, to save or delete documents from the bookmark panel. Between each test query the participants were asked to answer a fixed set of questions related to the retrieval system.		Relate assessments to task situations through a model of a semi-realistic situation and domain description. 2 scenarios: Scenario A Monitoring news and translation services Scenario B Information specialist /Consultant
Number of users	12users divided in 3 groups	8 participants	8 participants	12users divided in 3 groups
Languages	2 (English/Swedish)	2 (English/Swedish)	1 (French/French)	2 (English/Swedish)

System	Swedish: Siteseeker English: Inquery	The Clarity system for Swedish/English	2 systems: A text retrieval engine based on a standard retrieval System developed at SICS – with and without query expansion	Swedish: Siteseeker English: Inquery
Data	Swedish: Swedish TT Telegrambyrå CLEF corpus. English: LA Times from the CLEF corpus	Swedish: Swedish TT Telegrambyrå CLEF corpus. English: LA Times from the CLEF corpus	French texts from the CLEF corpus	Swedish: Swedish TT Telegrambyrå CLEF corpus. English: LA Times from the CLEF corpus
Inspection	Canned and ranked lists of searches outcomes of length between one and two hundred lines	Searches were performed in Swedish and retrieved results in either Swedish/English	No translation service was provided. Searches were performed in French, and retrieved results in French, Subjects were primarily Swedish speakers and moderately competent in the target language.	Canned and ranked lists of searches outcomes of length between one and two hundred lines
Presentations	Canned and ranked lists in Swedish/English	Ranked lists with news item headlines visible in original language. Bookmark panel.	Search window and then displays search results in a standard *ranked list*. *Document display window* next to the ranked list. Items from this list can be clicked and are then again displayed in the document display window for review. Bookmarked items can be *checked.* The highlighted text snippet is copied to the *answer window* and the document title automatically copied to the *reference window*.	Canned and ranked lists in Swedish/English
Relevance assessment	4 categories: - "not relevant" - "somewhat relevant" - "relevant" - "don't know".	"Inspected" "Saved in panel" "Saved as retrieval set"	4 categories: "Save Bookmark" A portion of text can be selected in document and saved. "Final Answer" "Uncertain Answer"	4 categories: - "not relevant" - "somewhat relevant" - "relevant" - "don't know".
Questionnaires	Pre- and post and after each query	Pre- and post and after each query	Pre- an post Questionnaire. A questionnaire designed for each of the two systems tested (with and without query expansion) and a final questionnaire.	Pre- and post and after each query

Outcome	1. Foreign-language texts took longer to assess and were assessed less well. 2. Task focus may have an effect on assessment performance 3. Methodology: Enhanced methodology for performing context-sensitive studies 4. Relevance assessment: We extended the binary relevance pairs with 2 more relevance levels. Relevance judgement aspect related to the - task domain - task given to the participant - topic of the query	1. Native (Swedish documents often used for inspection than foreign documents 2. Foreign (English) documents were discarded from the bookmark more often. Less confident 1st impression of foreign language? 3. Searches is native documents were reformulated more often. Recycling terminology from the target set?	1.Thesaurus and Term Expansion: The thesaurus was not useful due to its limited coverage; Subjects were frustrated by its inherent unpredictability and its patchy coverage. 2. Bookmark Panel: Users were happy about the bookmark panel and commented on it favourably. 3. Cooperation Users actually did cooperate and collaborate. The subjects communicated around 3 categories: - Topic - Search strategies - The language (Vocabularies and Translation) - System functionalities	*A: Relevance assessments* Relevance assessment takes longer in a foreign language than in a user's first language *B: Scenario effects on assessment effort* The simulated scenarios did make a difference for the assessment process, both by - measured access time and by - self-report by subjects *C. Topic effects on assessment effort* No effects on results by traditional relevance ranking were detectable.

3 Future directions

By no means does this report give an exhaustive inventory of the research questions opened by the systematic study of users, usage, contexts, and tasks in cross-lingual information access. We have in the studies we have performed shown that the effects we hypothesised were detectable, and in many cases stronger than we expected it to be. We found that measures such as effort expended on interpreting the results from a virtual search engine gave us purchase to separate the conditions we were studying.

This promises to be the starting points for a larger program of study: the task space, the possible use cases, the conceivable methodologies have not yet been explored in detail.

The contributions we offer this fare are methodological, in that our studies have not focussed on retrieval results per se, but on the effort users expend to understand what they retrieve – which effort we find is considerably larger and thus better measurable than in the monolingual case. We developed and used an extended simulated work task framework description: the Simulated Domain and Work-task Scenario (SDWS). This framework was enhanced with a domain description in addition to the work task description. Finally, we also find that this effort is larger than what users themselves would expect it to be, which has effects for any informed framing of policies for a multilingual environment.

References

Gonzalo, J. and Oard, D. (2003). The CLEF 2002 Interactive Track. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (eds), *Advances in Cross-Language Information Retrieval*. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 2002. Lecture Notes in Computer Science 2785. Springer 2003. pp 372-382.

Appendix A

The following is a full version of a Simulated Domain and Work-task Scenario (SDWS) (translated from the Swedish original) for I-clef query C053

General descriptions:

Domain:	Monitoring news and translation services
Work task:	Among your daily work-tasks you monitor and translate news information within a specific areas based on profiles set up by external customers. Your customers are usually companies and public institutions.

Situational description

Topic:	Genes and Diseases
Search task:	You have been assigned to monitor incoming news items that describe genes, which cause disease on humans. The customer especially wants documents that identify or report the discovery of a gene that is the source of any type of disease, syndrome, behavioural or developmental disorder in humans. Any information or document that reports the discovery of a defective gene that causes problems in humans is relevant. Documents that describe diseases and disorders caused by the absence of a gene are not relevant